Mixed Reality and Egocentric AI-Assisted Visualisation in Obstetric Ultrasound

Manuel Birlo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

November 5, 2025

Originality Declaration

I, Manuel Birlo confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

London, November 2025

Manuel Birlo

Impact Statement

This thesis explores new approaches to advancing Augmented Reality (AR) in medical contexts, with a focus on Mixed Reality (MR) systems. By blending real and virtual environments, MR holds significant potential in medicine, though challenges remain in clinical adoption. This thesis addresses some of these challenges through methods that leverage 3D virtual object visualisations and egocentric motion data from Optical See-Through Head-Mounted Displays (OST-HMDs). It also presents a scalable synthetic image generation method for training Deep Learning (DL) models in 3D hand and tool pose estimation, with a specific focus on obstetric sonography training.

Academically, this work contributes to human factors in surgical AR, MR-based medical education, and domain-specific generative AI. A collaboration with a UCL clinician lays the groundwork for further research. Publicly available datasets and source code support reproducibility, while the synthetic multimodal image generation method establishes a foundation for DL training in contexts with limited annotated data. Additionally, it highlights the potential of Reinforcement Learning (RL)-based physics engines to generate realistic hand grasp motions, further enhancing synthetic dataset realism.

Beyond academia, these methods could inform the development of commercial MR-based medical training simulators integrating AI-driven decision support, particularly benefiting midwifery and sonography programs in resource-constrained settings. The synthetic data generation approach could also support scalable dataset creation for AI model training in healthcare and other fields, where real data is scarce.

Overall, this thesis bridges AR, AI, and medical education, establishing a basis for future research and practical applications aimed at enhancing medical training and clinical practice through advanced technology.

Abstract

Augmented Reality (AR), particularly in the form of Mixed Reality (MR), holds strong potential to enhance clinical and surgical workflows. However, despite its promise and increasing research interest, demonstrated clinical utility of MR is rare. This thesis investigates specific factors that could contribute to more effective and intuitive AR systems in the future, with the overall goal of supporting real-time AR-assisted decision-making and procedural skill development.

First, a systematic literature review of 91 peer-reviewed publications on OST-HMD-based surgical applications was conducted. The review revealed that key technical limitations and human factors remain major barriers to widespread clinical adoption and proposed strategic research directions for improved study design and evaluation. Based on these findings, the focus turned to the underexplored area of AR-assisted training in obstetric ultrasound (US). A novel MR training application (CAL-Tutor) was developed to provide real-time 3D virtual overlay guidance for US probe positioning toward standardised fetal target US planes. A pilot user study demonstrated improved probe navigation and spatial understanding for novel users.

Findings from CAL-Tutor led to the development of a new method for ego-centric, markerless 3D hand and tool pose estimation. This approach leveraged a scalable synthetic data generation pipeline, combining a generative deep learning model for grasp synthesis with plausible 3D computer graphics-based rendering. The resulting multi-view, multi-modal dataset, HUP-3D, is the first of its kind in the research community. It includes over 31,000 synthetic samples and achieved state-of-the-art single-modality (RGB) accuracy (8.65 mm MPJPE) using the HOPE-Net pose estimation model. A dataset extension (HUP-3D-v2) and initial multi-modal

(RGB-D) evaluation finalise the main technical contributions of the thesis.

Overall, this thesis contributes to the design of reproducible, scalable and data-driven MR applications for clinical education and interactive procedural training, offering insights for future research at the intersection of MR, egocentric computer vision, and generative deep learning. The findings aim to pave the way for broader adoption of immersive AR- and AI-driven technologies in clinical education and procedural guidance.

Acknowledgements

As this journey comes to an end—one that allowed me to delve into interesting research, meet inspiring and kind people, and visit creative and exciting places—it also proved to be the most challenging project I have ever undertaken, and a true test of personal resilience during a time filled with difficult life circumstances and sacrifices beyond that I initially imagined. I'd like to take this opportunity to acknowledge a few people whose support was invaluable.

First, I thank my supervision team—Prof. Danail Stoyanov, Prof. Matthew J. Clarkson, and Dr. Philip J. "Eddie" Edwards—for trusting me to complete a PhD under challenging personal conditions, including part-time work as a professional software engineer, family duties, and remote collaboration. Without Prof. Stoyanov's empathetic and clear guidance, which I always trusted, I would certainly not be at this point. Prof. Clarkson supported me with professional advice and helped me refocus when I drifted off target. Dr. Edwards provided consistent technical guidance and collaboration throughout, giving me the confidence to reach milestones, even when the road became bumpier than expected. I would also like to thank Dr. Razvan Caramalau for his technical advice and collaboration.

It goes without saying that one's journey can only progress so far without the support of a great family. My mother, Hannelore Birlo, raised four children mostly on her own and on a limited budget—while also facing cancer at one point—a tremendous achievement that no PhD acknowledgement can truly do justice. My three siblings kept my competitive spirit alive, usually with good humour: Michael Birlo, a physicist working on remarkable optics projects in industry; Mitja Birlo, who followed his passion for cooking and is now a renowned haute cuisine chef;

and Stella Birlo, a trained geoscientist now in geoinformatics—both Michael and Stella began PhDs but ultimately chose careers in industry. My stepfather, German Schild—a business consultant and former managing director—has supported the entire family with consistent advice and care.

Sadly, some beloved family members are no longer with us, but their influence remains deeply present. My grandparents, Engelbert and Lilli Birlo—the most loving and caring grandparents one could wish for—played a major role in my upbringing and shaped much of who I am today. My great-cousin Giselbert Jäger was always there to talk to, though I regret not reaching out more often. And my aunt Milli Birlo consistently believed in me and provided encouragement during moments of self-doubt. All of them are deeply missed. I will do my best to honour their legacy, kind hearts, and the values they passed on.

Finally, completing a PhD while managing professional and family responsibilities requires sacrifice, commitment, and careful time management. Caring for my two beloved children, Raphael and Marcel Birlo, has been a privilege—though balancing fatherhood with the demands of doctoral research has not always been easy. My circumstances changed significantly over the course of this journey—not always for the better—but I have learned that even the most difficult goals can be achieved with passion and perseverance. I truly believe that with hard work and dedication, any personal goal is within reach, regardless of the circumstances.

Contents

1 Introduction 19
1.1 Motivation: Augmented Reality in Medicine 21
1.2 Background: Augmented Reality in Medicine 23
1.2.1 Definition and Motivation 23
1.2.2 Brief History 23
1.2.3 Challenges and Limitations in Medical AR 27
1.3 Research Objectives and Contributions 30
Goals 30
Contributions 30
1.4 Structure of the Thesis 32
1.5 List of Publications and Open-Source Contributions by the Author . 34

2 Assessing the Utility of OST-HMDs in AR-Assisted Surgery: A Systematic Review 37
2.1 Introduction 37
2.2 Background on OST-HMDs in Surgical AR 40
2.2.1 Safety of OST-HMDs in Surgical AR 41
Perceptual accuracy 42
Human-factor challenges 42
Addressed and unresolved challenges 43
2.2.2 Efficacy of OST-HMDs in Surgical AR 43
2.3 Methods 44
2.3.1 Literature Search and Scope 45

2.3.1.1 Contextual Review Literature 46
2.4 Literature Analysis Strategy 48
2.5 Analysis of the Literature Search 50
2.5.1 Annual Distribution of Selected Articles 51
2.5.2 Surgical Speciality 52
2.5.3 Device Type 52
2.5.4 Surgical Application Context 54
2.5.4.1 Surgical Guidance 56
2.5.4.2 Other Surgical Application Contexts 61
2.5.4.3 AR Visualisations 64
2.5.5 Validation and Evaluation of AR Systems 66
Evaluation methods 66
Experimental setting 67
2.5.6 Registration and Tracking in Surgical AR 69
2.5.6.1 Manual Alignment 70
2.5.6.2 Point-Based Registration 72
2.5.6.3 Surface Registration 73
2.5.6.4 Other Registration Methods 75
2.5.6.5 Tracking 75
2.5.7 Human Factors 77
2.5.7.1 Human Factors in AR 78
2.5.7.2 HCI Design Considerations in OST-HMD-Assisted Surgery 79
2.5.7.3 Human Factors Identification 80
2.5.7.4 Human Factors in Conventional Surgery Addressed by OST-HMDs 82
2.5.7.5 Persistent Human Factors of Proposed OST-HMD Solutions 85
2.6 Discussion 89
2.6.1 Classification of Human Factors 90

2.6.2 Potential Machine Learning Applications 91
2.7 Summary and Conclusions 92

3 CAL-Tutor: Investigating the Utility of a Mixed Reality Platform for Training in Obstetric Sonography 95
3.1 Introduction 95
3.1.1 Importance of Obstetric Ultrasound 96
3.1.2 Educational Challenges in Obstetric Ultrasound 97
3.1.3 Motivation for Using MR 98
3.1.4 Motivation for the CAL-Tutor System 99
3.2 Background and State of the Art 100
3.2.1 Tracking Methods in Medical AR 101
3.2.2 AR-Assisted Ultrasound Training 103
3.2.3 DL-Based Standard Plane Navigation Methods 105
3.3 Proposed Method 108
3.3.1 Design of the MR Concept 110
US probe tracking 111
3D virtual guidance during standard plane navigation 112
3.3.2 User Workflow 114
3.3.2.1 Manual Registration of the Baby Model 115
Optional semi-automated registration 115
3.3.2.2 Standard Plane Definition 117
3.3.2.3 Trainee Navigation to Standard Plane 119
3.3.3 User Data Recording 120
3.3.4 User Study 121
3.4 Results 123
3.4.1 Workload Assessment 124
3.4.2 Product Assessment 125
3.4.3 HoloLens 2 User Motion Data 126
3.5 Discussion 128
3.6 Conclusions and Future Work 129

3.6.1 Shift in Research Focus due to Challenges in Tracking Accuracy . . . 131

4 HUP-3D: A 3D Multi-View Synthetic Dataset for Assisted-Egocentric Hand-Ultrasound-Probe Pose Estimation 133

4.1 Introduction and Background . . . 133

4.1.1 Technical Foundations of Markerless Pose Estimation . . . 135

4.1.2 Synthetic Data Generation for Training Generative Pose Estimation Models . . . 136

4.1.3 Challenges and Limitations of Markerless Pose Estimation . 137

4.1.4 Contributions . . . 142

4.2 Method . . . 144

4.2.1 Initial Grasp Generation Using GraspIt!—Evaluation and Limitations . . . 145

4.2.2 Grasp Generation Using a Generative Model . . . 148

4.2.3 Grasp Rendering . . . 149

Camera view-angle sphere concept. . . . 150

Dataset comparison . . . 152

Noise characteristics . . . 152

4.3 Experiment: Pose Estimation Evaluation . . . 153

4.4 Discussion . . . 156

Dataset extension and multimodal evaluation . . . 157

4.5 Conclusion and Future Work . . . 158

5 HUP-3D-v2: Dataset Extension and Multi-Modal Evaluation for Hand–Probe Pose Estimation 160

5.1 Introduction . . . 160

5.1.1 Context and Technological Motivation . . . 161

Contribution of this chapter . . . 162

5.2 Method . . . 163

5.2.1 Application of HOPE-Net for 3D Pose Estimation . . . 165

5.3 Experiment: Pose Estimation Evaluation 169
  5.3.1 Single-Modality Evaluation 170
  5.3.2 Multi-Modality Evaluation 170
5.4 Discussion 173
5.5 Conclusion and Future Work 174
  5.5.1 Extending Data Diversity of HUP-3D Through Simulated Grasp Movements 176

6 Conclusions and Future Work 182
6.1 Summary of Contributions and Limitations 182
  6.1.1 Systematic Review on AR in Surgical Applications 183
  6.1.2 Mixed Reality for Medical Training 184
  6.1.3 Synthetic Dataset and Single-Modality Evaluation for 3D Hand–Probe Pose Estimation 185
  6.1.4 Dataset Extension and Multi-Modal Evaluation for 3D Hand–Probe Pose Estimation 186
6.2 Future Work 187
  6.2.1 Further Development of the CAL-Tutor MR Application . . 188
6.3 Reflections and Broader Impact 190
6.4 Final Conclusions 193

Appendices 194

A Additional Details on the Literature Review of Chapter 2 194

B Implementation Details of the Synthetic Image Dataset Generation Pipeline in Chapter 4 204
B.1 Explanation of Camera Sphere Parameters 205
B.2 Grasp Generation and Rendering Pipeline Configuration 206

Bibliography 208

List of Figures

1.1 The reality-virtuality continuum 21
1.2 The Medical Augmented Reality Framework by Navab et al. (2022) 24
1.3 First AR-assisted surgical guidance system by Kelly et al. (1982) . . 25
1.4 First use of HMD for obstetric AR by Bajura et al. (1992) 26
1.5 Varioscope AR: Early OST-HMD design in medicine 26
1.6 Roadmap of the thesis 32

2.1 Distribution of AR HMD-related surgical articles over 24 years . . . 39
2.2 Google Glass in Surgical Education Setup 41
2.3 Microsoft HoloLens hardware and surgical use 42
2.4 Systematic review search strategy 48
2.5 Systematic review results overview 51
2.6 Graphical illustration of included articles grouped by surgical speciality 54
2.7 Distribution of included papers by surgical speciality 55
2.8 Annual Distribution of articles by OST-HMD device from 2014-2020 55
2.9 Distribution of articles by **surgical application context** 56
2.10 Distribution of surgical guidance applications 57
2.11 Guided screw insertion and needle insertion examples 57
2.12 AR guidance for robotic surgery and endoscopy 59
2.13 AR applications in reconstructive and plastic surgery 60
2.14 Surgical training example application 61
2.15 Telementoring applications 62

2.16 Surgical anatomy assessment and teleconsultation applications in visceral surgery . . . 62
2.17 Distribution of included articles by type of AR visualisation 65
2.18 Experimental setting, from phantom to animal to clinical studies . . 67
2.19 Accuracy verification experiment examples using optical trackers . . 71
2.20 Point-based and surface registration examples . . . 73
2.21 Registration accuracy verification using a sheet of millimeter paper . 74
2.22 Human factors alleviated by OST-HMD AR . . . 82
2.23 Distribution of persistent human factors of the proposed AR surgical approaches . . . 86

3.1 Standard ultrasound planes, used for fetal biometry estimation . . . 97
3.3 Potential areas to consider for innovation in future US training . . . 100
3.2 CAL-Tutor system setup . . . 100
3.4 Overview of an Mixed Reality (MR)-based obstetric pulsed-wave Doppler Ultrasound (US) training system . . . 105
3.5 Overview of the proposed DL-assisted US system by Droste et al. (2022) . . . 107
3.6 Core components of CAL-Tutor's Unity scene . . . 113
3.7 Unity concepts of ultrasound probe navigation . . . 114
3.8 Illustration of the CAL-Tutor's user workflow phases . . . 116
3.9 CAL-Tutor 3D virtual object alignment options . . . 118
3.10 NASA-TLX workload assessment result . . . 125
3.11 Product assessment result . . . 127
3.12 Visual attention profiles of study participants . . . 127

4.1 Selected qualitative results of hand and tool pose estimation 138
4.2 Illustration of hand articulation with high degrees of freedom 139
4.3 Egocentric images showing hand-object interactions in everyday kitchen scenarios . . . 139

4.4 RGB-D video sample frames illustrating hand pose ground truth annotation using sensors 140
4.5 Sample images from a synthetic hand-object dataset 142
4.6 Grasp Generation and Rendering Pipeline Overview 144
4.7 Sample hand grasp configurations of a GraspIt! Eigengrasp planner hand motion sequence 147
4.8 Sample manual auto grasp using GraspIt! 147
4.9 Grasp conversion and sphere camera setup 149
4.10 Sample frames from the HUP-3D dataset 151
4.11 Qualitative results, shown with 4 test images from HUP-3D 156

5.1 Interpenetration of Clarius probe and fingers 164
5.2 HOPE-Net architecture for single-modality input using RGB images only 165
5.3 HOPE-Net architecture for multi-modality input using RGB and depth images 168
5.4 Qualitative results, shown with 4 test images from HUP-3D-v2 . . . 171
5.5 Schematic illustration of a potential future grasp rendering pipeline . 179
5.6 Synthetic RGB example image of a hand with mismapped hand pose 180

List of Tables

2.1 Screening Criteria and Extracted Data from the Literature Review . 49
2.2 Distribution of the included articles per surgical speciality 53
2.3 Papers by tracking method 76
2.4 Identified Human Factors, grouped into 3 categories 81

3.1 Hardware used in the implementation of the CAL-Tutor system . . . 109
3.2 Software used in the implementation of the CAL-Tutor system . . . 109
3.3 Data recorded by the CAL-Tutor application 121
3.4 The two experimental conditions of the user study 123
3.5 NASA-TLX questionnaire results for the six study participants . . . 124
3.6 Individual product assessment result of all six study participants . . 126

4.1 Pose estimation dataset comparison 152
4.2 HOPE-Net training hyperparameters. 153
4.3 MPJPE (mm) comparison for pose predictions on clinical egocentric datasets 155

5.1 Performance of HOPE-Net on different datasets and configurations . 171

A1 Acronyms used in Table A2 (surgical context) 195
A2 Studies listed by OST-HMD, surgical context and surgical procedure 195
A3 Addressed and Persistent Human Factors 197
A4 Description of AR visualisation, conducted experiments and accuracy 199

B1 Frame generation configuration 207

List of Abbreviations

AI Artificial Intelligence. 19, 31, 35, 101, 177, 190

AR Augmented Reality. 19–27, 29–32, 34–41, 43–50, 52, 58–61, 63–67, 69–71, 77–79, 81, 82, 84–91, 93, 95, 98–104, 111, 119, 129, 175, 183, 184, 188, 190, 191, 193, 194, 199–203

AV Augmented Virtuality. 20

CNN Convolutional Neural Network. 106, 107, 135, 154

CT Computed Tomography. 23–25, 63, 64, 72–74, 88, 97, 104, 199, 201–203

CTA Computed Tomography Angiography. 60

DL Deep Learning. 29, 31, 33, 35, 101, 105–108, 120, 130, 131, 134, 135, 137, 145, 148, 153, 155, 156, 159, 160, 164, 165, 167, 177, 178, 182, 183, 187, 193

HCI Human–Computer Interaction. 78–81, 85, 88–92

HMD Head Mounted Display. 19, 20, 24, 26, 38–40, 43, 46, 47, 51, 91, 92

IMU Inertial Measurement Unit. 106, 107

MIS Minimally Invasive Surgery. 21, 78

ML Machine Learning. 19, 91, 92, 136, 138–142, 187, 188, 191–193

MPJPE Mean Per Joint Position Error. 154, 155, 170–172

MR Mixed Reality. 14, 20–22, 27, 29–36, 40, 43, 46, 57, 60, 91–96, 98, 99, 101, 104–106, 108–110, 120–126, 128–130, 133, 134, 158, 161, 175, 182–185, 187, 188, 190–193

MRI Magnetic Resonance Imaging. 61, 63, 64, 97, 109–111, 113, 203

NN Neural Network. 33, 92

OST-HMD Optical-See-Through Head-Mounted Display. 19–25, 27–30, 32, 33, 35–45, 47–59, 61, 66, 67, 69–71, 75–82, 84–93, 98, 102, 108, 109, 120, 129, 134, 135, 137, 161, 182–184, 190, 193–196, 203

RL Reinforcement Learning. 107, 177, 180, 192

RNN Recurrent Neural Network. 106

US Ultrasound. 14, 22, 24–26, 29–34, 36, 58, 60, 61, 95–115, 117–124, 126–131, 133–135, 143, 144, 146–148, 150, 152, 154, 155, 157–162, 164, 165, 170, 173–177, 179, 181, 182, 184–189, 191–193

VAC vergence-accommodation conflict. 28

VC Virtuality Continuum. 20

VR Virtual Reality. 19–21, 46, 47, 101, 104

VST Video See-Through. 38, 39

VST-HMD Video See-Through Head-Mounted Display. 19, 20, 23–26, 29, 38, 53, 54

Chapter 1

Introduction

Unlike Virtual Reality (VR), where the user is fully immersed in a computer-generated world and can no longer see the real environment, Augmented Reality (AR) allows the user to view the real world while simultaneously perceiving superimposed virtual objects. In this sense, AR has the potential to enhance the user's ability to perceive and interact with the real world [1]. AR is widely seen as a potentially transformative technology in several domains, such as engineering, urban planning and healthcare [2]. The term AR was first introduced in 1992 in the context of an Optical-See-Through Head-Mounted Display (OST-HMD) used for aircraft manufacturing. However, the first conceptual prototype of an AR system had already been built in 1968 in a pioneering work by Sutherland, who presented the first Head Mounted Display (HMD) [3]. Even though the term "AR" had not yet been introduced, Sutherland's prototype can be considered an early example of what is now referred to as an OST-HMD, a HMD that allows users to perceive 3D-rendered content while still seeing the real world. Today, other types of AR are being used, such as Video See-Through Head-Mounted Display (VST-HMD), allowing to see the real world only through a live camera feed, with overlaid AR content. Or projector-based AR that uses projectors to overlay digital content onto physical subjects or objects. AR is widely regarded as a key component of the Fourth Industrial Revolution[1] [4], next to advances in fields such as Artificial Intelligence (AI), Machine Learning (ML) and cloud computing. However, despite

---

[1] Also known as ”Industry 4.0”. This term was first introduced in 2011, but gained significant momentum in 2016 as related technologies began to be more widely adopted across industries.

its transformative potential and growing interest in AR in various industries, its integration into routine processes presents several challenges, including technical limitations and user acceptance [5]. These challenges are particularly significant in the medical field, where AR has the potential to revolutionise diagnostics [6], surgery [7], and medical training [8], but faces hurdles in clinical adoption.

This thesis explores challenges and potential solutions in the area of medical AR. More specifically, the initial phase investigates the potential of OST-HMDs, a specific type of AR device, in surgical applications. OST-HMDs are a category of wearable AR HMDs that preserve the user's egocentric view of the world and do not alter the user's viewpoint or field of view while displaying overlaid 3D-rendered content [9]. In addition, OST-HMDs allow users to see the real world directly, in contrast to another type of HMD called VST-HMD, in which the user's egocentric field of view is perceived through cameras.

Some OST-HMD devices like the Microsoft HoloLens (Microsoft Corporation, Redmond, USA) and its successor, the HoloLens 2[2], enable users to interact with virtual objects, which in turn can also interact with physical objects, independently of user manipulation. Such applications expand the classical concept of AR to a technology called MR. A popular definition of MR can be found in [10], where a Virtuality Continuum (VC) has been defined (Fig. 1.1). VC is defined as a spectrum ranging from reality to VR, with AR and Augmented Virtuality (AV) in between. According to [10], AV describes the part of the spectrum where real-world elements are integrated into a predominantly virtual environment. Consequently, along the VC, MR resides as a separate spectrum between AR and AV. In this sense, MR can be defined as a technology that enables real and virtual elements to interact and merge into one another. While these concepts illustrate the theoretical evolution from AR to MR, practical adoption of MR has proven more complex. Having noted the controversy over the transformative potential of AR and its barriers to widespread adoption, the recently announced discontinuation [11, 12] of Microsoft HoloLens reflects this disparity.

---

[2] `https://www.microsoft.com/en-us/hololens`, accessed on 17 September 2024

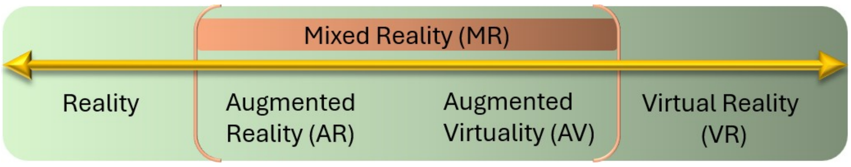


Figure 1.1: The reality-virtuality continuum, which extends from reality to VR, from left to right. Figure adapted from Milgram et al. [10].

Despite the conceptual distinction between the broader term AR and the more specific term MR, this thesis will use them interchangeably during the implementation and experimental validation stages, beginning in Chapter 3, due to the increasing popularity, technological advancements, and widespread adoption of MR-based OST-HMD devices, such as the HoloLens 2, within the research community. A more detailed presentation of the motivation behind this thesis is presented in section 1.1. Section 1.3 presents the goals and corresponding contributions of this thesis.

1.1 Motivation: Augmented Reality in Medicine

The history of medicine has shown that it takes time to establish a novel technology aimed at replacing conventional procedures and thus improving patient care. As a core medical field that impacts patient health and safety, surgery has always been a vital medical field that has benefitted from such innovations. The transition from open surgery to Minimally Invasive Surgery (MIS) marked a major breakthrough across surgical specialties, significantly improving patient outcomes by reducing blood loss, hospitalisation time [13], and the risk of surgical site infections [14, 15], while also lowering hospital costs [16]. Among the various surgical techniques classified under MIS are laparoscopic surgery, endoscopic surgery, and robotic-assisted surgery, each representing a significant milestone in the evolution of modern surgery. While MIS is well established in modern surgical practice, AR-assisted surgery – considered a complementary innovation aimed at enhancing routine procedures – remains in its early stages of adoption despite having been proposed several decades ago. Among the various applications of AR in surgery, AR-assisted surgical guidance has emerged as one of the most prominent. The concept was initially introduced in the 1980s [17, 18], primarily focusing on overlaying anatomical data to assist in surgical navigation.

Since those early implementations, subsequent research has explored additional potential benefits of AR, including improved visual perception, ergonomics, hand-eye coordination, safety, reliability, and repeatability, ultimately aiming to enhance surgical outcomes. However, despite over four decades of research, the promise of AR has yet to be fully realised in routine clinical practice. In recent years, likely driven by technological advancements such as the release of the first Microsoft HoloLens OST-HMD in 2016, a significant increase in related research publications [19] can be observed. However, despite promising preliminary results, sufficient evidence supporting the widespread adoption of AR-assisted surgery in routine clinical practice is still lacking [20, 21, 19, 22, 23]. With the current upward trend in surgical AR research and the persistent challenges that hinder its clinical adoption—such as those related to hardware, registration and perceptual accuracy, and user acceptance, discussed further in Section 1.2.3—this thesis focuses on three main areas of work, particularly in the context of medical OST-HMDs:

First, after analysing proposed solutions and limitations of surgical OST-HMD-based AR, the focus shifts to AR for medical training. Obstetric sonography emerges as a promising yet underexplored field in which HoloLens 2-based MR could enhance clinician training. Second, this thesis addresses a key challenge in conventional obstetric sonography training: a clinician's ability to mentally map 2D ultrasound images to the 3D anatomy of the fetus. An MR-based approach is introduced to enhance this aspect of clinical training. Finally, this thesis delves deeper into the technical aspects to meet a core requirement of the HoloLens 2 application and explores egocentric computer vision-based tracking of a user's hand and US probe. Prior to outlining the goals and contributions of this thesis in Section 1.3, Section 1.2 presents a background on AR in medicine, covering its historical development and key challenges.

1.2 Background: Augmented Reality in Medicine

1.2.1 Definition and Motivation

Medical AR refers to the application of AR in a medical context, with the aim of improving conventional workflows and ultimately providing additional benefits to patients. Its underlying motivation stems from the need to visualise data alongside real-world medical content—such as training devices, medical hardware, and patients—within the same physical space [24], which facilitates more effective integration of visual information, and hence more effective medical workflows, ultimately benefiting patients. Three primary types of AR devices are being used in medicine (as already described in Chapter 1): OST-HMD, VST-HMD and Projector-Based AR systems. Various application fields exist such as preoperative planning, intra-operative visual guidance, medical education and training, rehabilitation, diagnostic imaging, oncology and telemedicine. Depending on the application field different types of AR devices are being used. While for example OST-HMDs and VST-HMDs are more common in preoperative planning, intra-operative guidance, medical education and training, telemedicine and diagnostic imaging, projector-based AR systems are rather used in rehabilitation [25, 26] and oncology [27]. Chapter 2 discusses these types of AR devices and the aforementioned surgical application fields in more detail. Before delving into the history of medical AR in the next Subsection 1.2.2, the reader is referred to an existing modern definition of medical AR provided by Navab et al. [28]. It includes the definition of a Medical AR Framework (Fig. 1.2), which describes an interconnected relationship between its primary components and illustrates the information flow between the physical world and the AR system.

1.2.2 Brief History

The first reported use of AR guidance in medicine dates back to 1982 in the context of neurosurgical navigation [17]. In this computer-monitored stereotactic neurosurgery system, a surgical microscope was rigidly attached to a stereotactic frame, allowing Computed Tomography (CT)-based tumor outlines to be overlaid directly

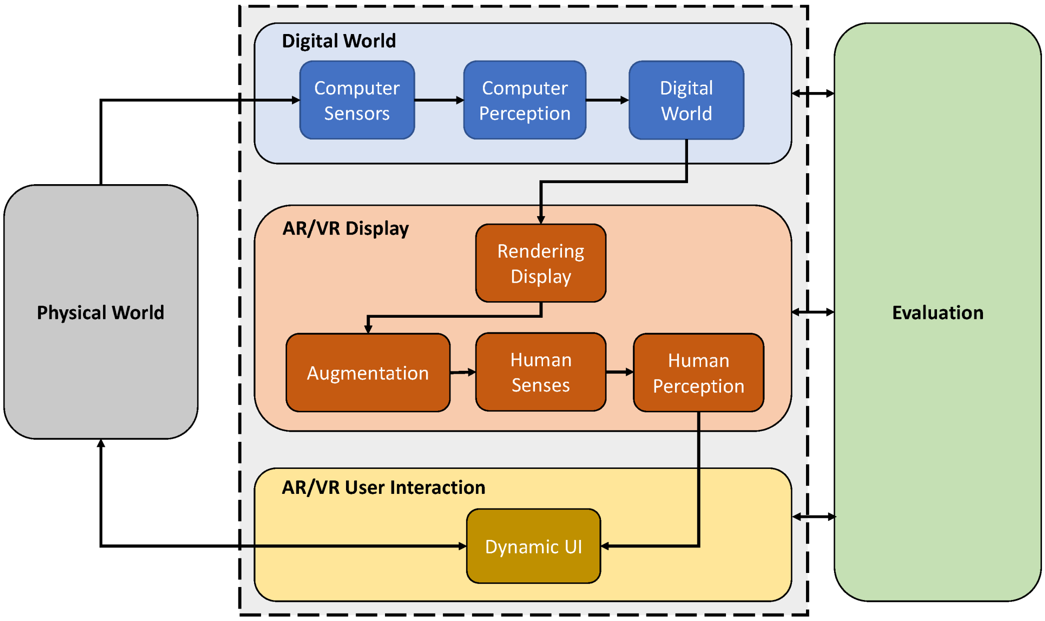


Figure 1.2: The Medical Augmented Reality Framework by Navab et al. (2022) illustrates the relationships between its four primary components: Digital World, AR Display, AR User Interaction, and Evaluation. Figure reused from Navab et al. [28].

into the microscope's view (Fig. 1.3). Four years later, in 1986, [18] expanded this method and projected CT images in a surgical microscope. A few years later, in the 1990s, early VST-HMD systems were introduced, approaching the first reported use of HMD-assisted medical AR: In 1992, within the field of obstetric sonography, the first system for medical imaging was introduced, overlaying real-time US images directly onto a pregnant woman [29] (Fig. 1.4a and 1.4b). In 1994 [30] presented one of the first reported instances of a system for real-time 3D US visualisation in obstetric sonography. A volume-reconstructed fetus, generated from US slice acquisitions, was overlaid onto its corresponding physical location within the mother's abdomen. In 1997 the first AR-assisted surgical navigation system for tumor resection in endoscopic ear, nose, face and throat (ENT) surgery was presented [31]. In this method, AR guidance was achieved through rectangles overlaid onto the live endoscopic video and pointing to the resection target.

OST-HMD systems were less common in medical AR applications than VST-HMD systems in the 1990s. This can likely be attributed to the technological lim-

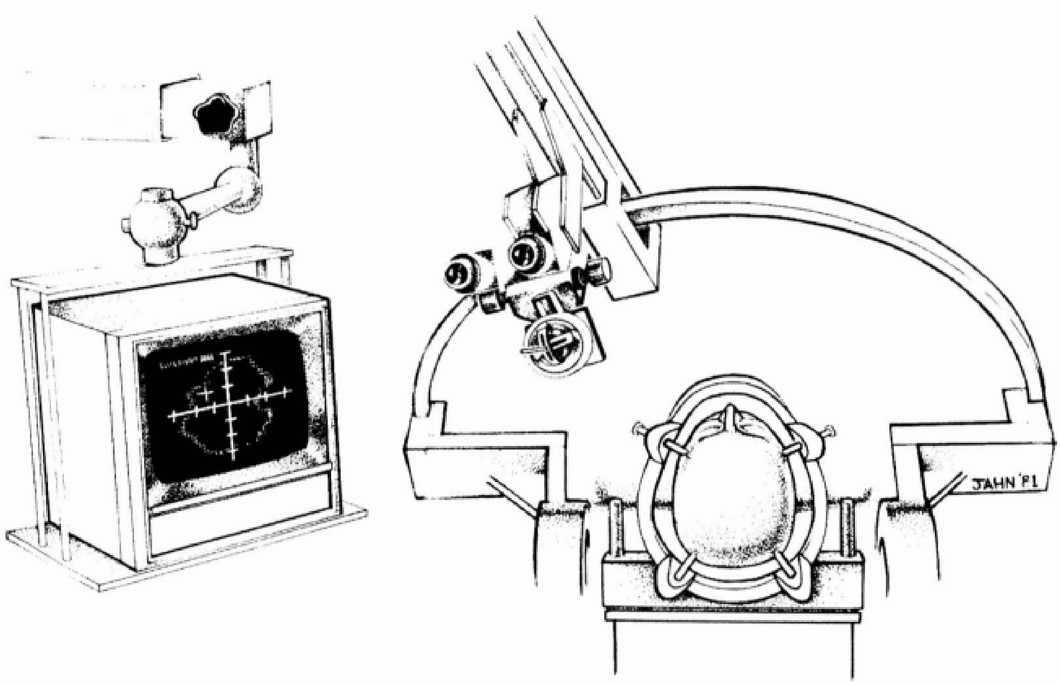
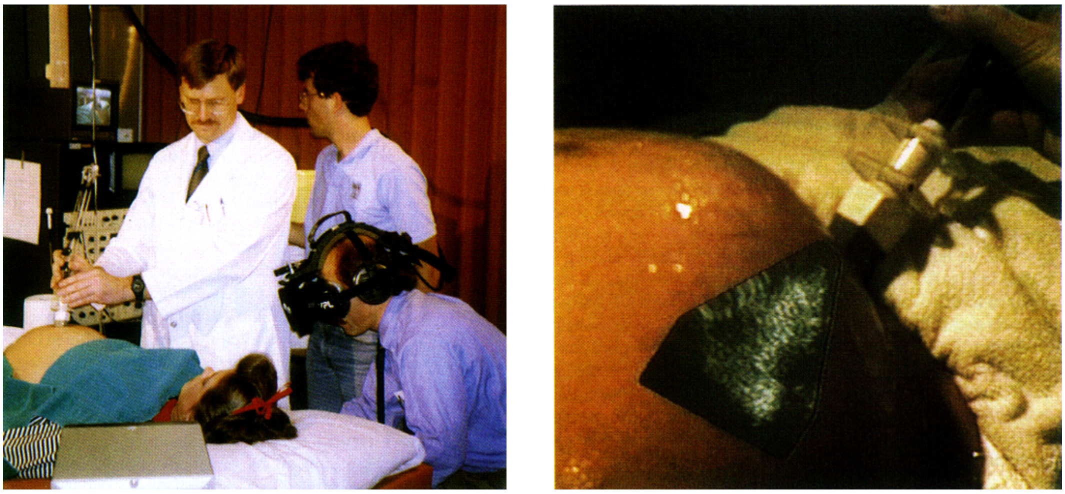


Figure 1.3: Illustration of the surgical system developed by Kelly et al. (1982), considered the first reported use of AR-assisted surgical guidance. Tumor outlines from preoperative CT scans were displayed on a computer monitor to assist navigation with a surgical microscope that was rigidly attached to a stereotactic frame. Figure reused from Kelly et al. [17].

itations of OST-HMD at that time, such as insufficient registration accuracy and a limited field of view. One of the first attempts to superimpose a 3D-rendered knee-joint visualisation on a leg model was demonstrated in [32]. In the 2000s, technological advancements continued with key contributions such as the development of a head-mounted operating microscope for computer-aided surgery in 2000 [33, 34], which can be categorised as OST-HMD (Fig. 1.5a and 1.5b). Based on preoperative CT data and a preoperative planning software, non-real-time OpenGL[3]-rendered navigation data, such as implants and a drill, was displayed to the user. In 2002 a pioneering VST-HMD-based US-guided needle biopsy application was proposed [35]. In a randomised, controlled trial the proposed AR-assisted method was compared against conventional US-guided needle biopsies. The distance of the biopsy to the ideal target position was measured, and results showed that the AR-assisted method led to smaller mean deviations from the target position.

---

[3] https://www.opengl.org/



(a) **(b)**

Figure 1.4: The first reported use of HMD-assisted medical AR by Bajura et al. (1992), within the field of obstetric sonography. (a) AR setup with US technician scanning a pregnant woman, and a person wearing a VST-HMD to observe live 2D US images slices in 3D. (b) A sample video frame from the VST-HMD user's left-eye perspective, showing a US slice registered to the US probe and superimposed onto the patient's abdomen. Figures reused from Bajura et al. [29]

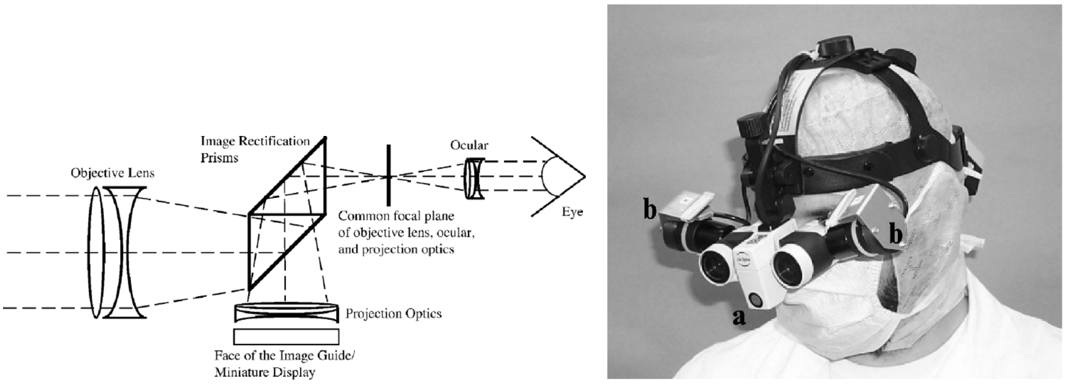


Image Rectification Prisms
Objective Lens
Ocular
Eye
Common focal plane of objective lens, ocular, and projection optics
Projection Optics
Face of the Image Guide/ Miniature Display
b
b
a

(a) **(b)**

Figure 1.5: An early design of an OST-HMD in medicine: The Varioscope AR. (a) The optics illustrating the concept of image overlay in the Varioscope AR: A miniature computer display projects an additional image into the focal plane of the Varioscope's objective lens. The combined image consisting of real world view and computer-generated content can be viewed through the ocular. (b): The Variscope AR prototype, consisting of a commercial Varioscope (Life Optics, Vienna, Austria) (marked "a"), and two miniature computer displays (marked "b"). Figures reused from Birkfellner et al. [34].

In 2004 a projector-based AR system for intra-operative navigation in maxillofacial surgery was presented, called ARSys-Tricorder [36]. The aim of this system was to meet the need for more precise, submillimeter-level superposition of preoperative skull transplant design onto the patient during surgery. Pre-operative planning information such as osteotomy lines were projected onto a semitransparent mirror that was positioned above the patient anatomy. However, an experimental evaluation of this method was missing. For readers interested in a comprehensive overview of medical AR advancements, including its historical development up to the 2000s, a detailed literature review is available in [24].

Moving forward to the 2010s, advances in medical AR were partially attributed to the release of popular OST-HMDs such as Google Glass (Google Inc., Mountain View, CA) in 2014 and Microsoft HoloLens (Microsoft Corporation, Redmond, USA) in 2016. Further details on relevant applications using Google Glass and Microsoft HoloLens will be discussed in Chapter 2. Readers who wish to get a more in-depth overview of MR and AR technology in medicine as of 2018 are referred to [37].

1.2.3 Challenges and Limitations in Medical AR

Besides the potential advantages of medical AR, as presented in subsection 1.2.1, there are several factors that prevent these new methods from replacing conventional routine practices. One of the main limitations is the lack of standardised evaluation methods for proposed AR-assisted medical solutions, which hinders their regulatory approval and adoption [38]. Focusing specifically on OST-HMDs, these devices face several technical limitations. One of the most significant limitations is the inaccurate spatial registration of virtual overlays onto real-world structures. Especially in patient safety critical applications like AR-assisted surgical guidance [39], where an accurate registration of virtual content with physical patient-anatomy or surgical tools is critical and ensures that 3D virtual objects[4] are constantly perceived at their correct physical location. However, achieving and maintaining sufficiently accurate 3D virtual object registration with OST-HMDs is technically challenging due to the underlying requirement for precise device calibration. Even minor calibration errors can result in substantial deviations, which are unacceptable in surgical

---

[4] While Microsoft refers to the 3D content rendered by OST-HMD devices such as the HoloLens 2 as “holograms”, this thesis uses the terms “3D virtual object”, “3D virtual content” and “3D virtual overlay” to describe the accurate technical nature of the visualisation.

guidance tasks [40].

Besides 3D virtual object registration accuracy, OST-HMDs also face hardware-related limitations. For example, the first version of the Microsoft HoloLens, a widely used OST-HMD released in 2016, attracted substantial research attention but suffered from a restricted field of view, insufficient battery life and a heavier weight [41]. There are also perceptual issues that limit the utility of OST-HMDs. A study published in 2020 suggested that the first version of the HoloLens is not suited for surgical guidance tasks [42] due to impaired perceptual accuracy. This problem can be attributed to the fact that the focal length of the HoloLens is approximately 2 meters. As a result, all virtual objects appear optically focused at this distance, regardless of their intended depth. Consequently, during manual tasks in which the hands are closer to the eyes than this focal length, the eyes cannot simultaneously focus on both virtual and real content. Another problem, as noted by [42], are the vergence-accommodation conflict (VAC) [43] in binocular vision and the poor alignment of virtual content to the scene due to failed per-user display calibration. The HoloLens 2 addressed some of the aforementioned issues like wider field of view, lower weight, and reduced focal length at around 1.5m. However, even at this reduced focal length a mismatched vergence accommodation for virtual content in the peri-personal space can occur [44].

Such technical limitations of OST-HMDs directly impact user acceptance when working with these devices in a medical context that requires precision and concentration. For example, the VAC can lead to visual discomfort and fatigue [45]. In this context, technical limitations are closely related to human factors [46], which, in this thesis, refers to the individual perceptual and physical characteristics of each user. These factors vary and must be considered when designing OST-HMDs and their respective medical applications. In addition, human factor limitations can also be independent of technological constraints and can simply be attributed to a user's individual ability or willingness to work with novel technologies. Although the definition of human factors is sometimes distinguished from perceptual issues [46], this thesis considers perceptual issues as part of human factors.

In that sense, example human factors, as defined in Chapter 2, are impaired depth perception, Individually different visual processing capabilities between dominant and non-dominant eye, perception of spatial relationships between real and virtual objects, confidence and frustration.

Taking into account the technological and human factor limitations of OST-HMDs, it is important to note that medical applications requiring high perceptual accuracy when overlaying virtual content onto real-world counterparts may face significant challenges [42]. One such prevalent application in the research community is OST-HMD-assisted surgical guidance. In particular, surgical specialties such as neurosurgery and orthopedic surgery, where rigid bone structures can serve as fixed reference points for 3D-rendered overlays, may significantly benefit from AR. Recent studies show that OST-HMDs for intraoperative navigation show promising results, but conclude that further work is required before widespread clinical adoption can be achieved [47, 44, 48]. A recent qualitative study investigated whether the superimposed display of virtual content is suitable for OST-HMD and VST-HMD surgical navigation in the context of widespread clinical adoption [49]. The authors concluded that such systems are not yet suitable for routine clinical use due to underlying human factor issues. Despite advancements in commercial OST-HMD technology, maintaining accurate spatial 3D virtual object registration remains a significant challenge in clinical applications where precision is crucial, such as surgical guidance [50, 40].

Motivated by these limitations, this thesis explores the current state of OST-HMD-assisted surgical AR through a systematic review, investigates the potential of MR for medical training via the CAL-Tutor—an MR prototype for obstetric US education—and proposes a novel markerless hand and tool pose estimation method to support egocentric MR interaction. The generation of synthetic images for state-of-the-art Deep Learning (DL)-based pose prediction models forms an essential component of the latter approach, and the specific aims and contributions of this work are outlined in Section 1.3.

1.3 Research Objectives and Contributions

This thesis contributes to the evolving field of AR-enabled technologies for clinical and educational use, aiming to advance both training methodologies and technical capabilities in MR environments. This section lists the primary research goals and summarises the corresponding technical contributions, which collectively address persistent limitations identified in surgical AR and propose novel methods for MR-driven medical training and interaction.

Goals. Building on the challenges outlined in Section 1.1, this thesis pursues the following objectives:

1. **Assess the current state of AR-assisted surgery:** Conduct a systematic review to identify opportunities and limitations in AR-assisted surgical systems and propose strategic directions for future research.
2. **Develop an MR training application for obstetric US:** Design and implement a targeted training tool using egocentric interaction and 3D-rendered guidance, translating insights from surgical AR into a medical education context. Obstetric US was chosen as the medical application domain due to its inherent perceptual and spatial challenges, making it a suitable context for evaluating the proposed MR-assisted methods.
3. **Advance egocentric hand and US probe tracking:** Propose and validate a markerless method for 3D hand-probe pose estimation in OST-HMD-based systems, addressing challenges in marker-based egocentric probe tracking.

Contributions. Following the previously mentioned goals of the thesis, several contributions were made that address persistent challenges in medical AR, particularly in the context of OST-HMD-based systems. Consequently, new insights, tools, and methods were delivered that aimed at improving both educational applications and technical capabilities in MR environments.

1. **Systematic literature review and research gap analysis:** A comprehensive systematic review of OST-HMD-based AR applications in surgery was

conducted. The review identified recurring technological and human factor limitations that hinder clinical adoption. Findings from this review informed the scope of this thesis, directing focus toward medical training applications.

2. **CAL-Tutor: An innovative MR training platform for obstetric US:** As the first technical contribution of this thesis, CAL-Tutor, a HoloLens 2-based MR platform, was developed to address the underexplored area of AR-assisted education in fetal US. A preliminary user study demonstrated the system's feasibility and highlighted potential for future research towards evaluation of user motion data for AI-assisted skill assessment.

3. **HUP-3D: A synthetic dataset with DL-based evaluation for markerless 3D hand-probe pose estimation** To address limitations in marker-based US probe tracking observed in CAL-Tutor, a novel markerless approach for joint 3D hand-tool pose estimation was developed. Central to this contribution is a scalable pipeline for generating synthetic images, depicting hand-probe grasps in an obstetric sonography setting, and intended for training state-of-the-art DL-based 3D pose estimation models. The resulting multi-view and multi-modal dataset, HUP-3D, achieved state-of-the-art DL-based 3D pose prediction performance on a clinical dataset.

4. **HUP-3D-v2: Extension of HUP-3D and multi-modal evaluation** In a final contribution, the HUP-3D dataset was extended to support a greater data diversity and DL model training generalisability when trained on this data. Grasp synthesis and multi-modal (RGB-D) evaluation highlighted key challenges, indicating areas for future improvement.

Collectively, these contributions support the broader goal of improving the usability, reliability, and applicability of MR in medical education, while also drawing from foundational challenges in AR-assisted surgery and addressing those that hinder its widespread adoption. To provide a clear roadmap for how these objectives are addressed, the structure of the thesis reflects a progression from in-depth problem analysis and conceptual design to implementation and technical validation. The

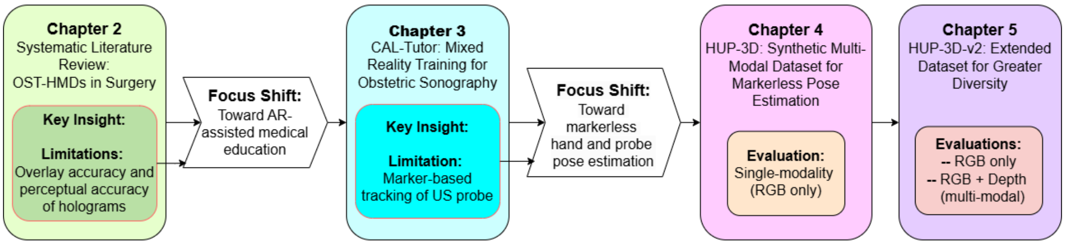

Chapter 2
Systematic Literature Review:
OST-HMDs in Surgery
Key Insight:
Limitations:
Overlay accuracy and perceptual accuracy of holograms
Focus Shift:
Toward AR-assisted medical education
Chapter 3
CAL-Tutor: Mixed Reality Training for Obstetric Sonography
Key Insight:
Limitation:
Marker-based tracking of US probe
Focus Shift:
Toward markerless hand and probe pose estimation
Chapter 4
HUP-3D: Synthetic Multi-Modal Dataset for Markerless Pose Estimation
Evaluation:
Single-modality (RGB only)
Chapter 5
HUP-3D-v2: Extended Dataset for Greater Diversity
Evaluations:
-- RGB only
-- RGB + Depth (multi-modal)

Figure 1.6: Roadmap of the thesis: From a systematic literature review of OST-HMD-assisted surgery to markerless Hand and US probe pose estimation.

following Section 1.4 outlines the content and purpose of each chapter to guide the reader through the remainder of the thesis.

1.4 Structure of the Thesis

The three main contributions of the thesis, as mentioned in the previous Section 1.3, led to a chapter-based thesis structure that is illustrated in Figure 1.6. Chapters 2 through 5 present the core technical contributions of the thesis. Associated peer-reviewed publications related to Chapters 2-4 are then listed in Section 1.5. While this Introduction chapter provides a broad overview of the medical AR landscape, specialised background on OST-HMDs applications in surgery, AR-assisted US guidance, and egocentric pose estimation is detailed in Chapters 2, 3, and 4, respectively. The subsequent chapters are structured as follows:

Chapter 2: Presents a systematic literature review of OST-HMD-assisted surgical applications, covering 91 studies published between 2013—the year Google Glass (Google, Inc.), then a widely known OST-HMD, was released—and 2020. Selected publications from 2021 to 2024 are referenced to illustrate ongoing trends, but are not included in the formal meta-analysis. Key technical and human factor limitations were identified, motivating a shift in focus from surgical to educational MR use cases. This shift was based on the premise that educational scenarios impose less stringent accuracy demands. The resulting work led to the development of the CAL-Tutor MR application for training in obstetric sonography, which is the focus of the next chapter.

Chapter 3: Introduces CAL-Tutor, a novel HoloLens 2-based MR application for obstetric US training, developed in response to the limitations of OST-HMD-assisted surgery identified in Chapter 2—particularly challenges related to overlay and perceptual accuracy of 3D-rendered content. Obstetric US training was selected as the target medical education context due to its inherent difficulty: trainees must mentally map 2D US images to 3D fetal anatomy, a cognitively demanding task that contributes to high inter-operator variability. To address such learning challenges, CAL-Tutor provides 3D virtual overlay guidance to direct the US probe toward predefined anatomical targets, thereby supporting the trainee's spatial awareness and potentially shortening the learning curve. A user study demonstrated system feasibility and highlighted limitations of marker-based US probe tracking, leading to the development of a markerless alternative.

Chapter 4: Proposes a novel approach for markerless 3D hand-tool pose estimation using synthetic data, motivated by the limitations of the marker-based US probe tracking method used in the CAL-Tutor MR application (Chapter 3). The chapter introduces HUP-3D, a synthetic, multi-modal dataset of hand-US probe grasps designed for training DL-based 3D pose estimation models. The dataset emphasises image diversity and includes RGB, depth, and segmentation mask frames, along with ground truth pose data inherently generated during the rendering process. Realistic hand grasp poses were generated using a pre-trained generative Neural Network (NN), followed by manual selection to ensure anatomical plausibility. A spherical camera sampling strategy was employed to capture synthetic hand-probe grasps from multiple perspectives, including both egocentric and third-person views. Evaluation using single-modality (RGB) input and state-of-the-art HOPE-Net [51] model demonstrated competitive results.

Chapter 5: Extends HUP-3D by introducing a second US probe to the dataset and performing a multi-modal (RGB-D) evaluation. Grasp synthesis using larger handheld objects, dataset generalisability, and the identification of challenges for future work are core elements of this chapter. Building upon the competitive performance of the purely synthetic HUP-3D dataset on keypoint localisation tasks

(Chapter 4), this chapter enhances image diversity and improves generalisability through the addition of a second US probe model in HUP-3D-v2. To evaluate its effectiveness, the same state-of-the-art model used previously was applied, and the dataset's multi-modal capabilities were tested by incorporating depth frames alongside RGB inputs. While this extension demonstrated technical feasibility of multi-modal, multi-probe configurations, initial pose prediction performance was lower than single-probe, single-modality results—highlighting the need for further optimisation.

Chapter 6: Summarises the main contributions of the thesis, emphasising how the developed systems—the CAL-Tutor MR application, the HUP-3D synthetic dataset generation and validation, and its extension into HUP-3D-v2—collectively advance AR-assisted training methods. It discusses the practical and technical limitations encountered, proposes target future research directions, and reflects on the broader implications for AR-assisted medical education and procedural guidance. The chapter closes with concise remarks summarising the lessons learned and the thesis's contribution to the field.

1.5 List of Publications and Open-Source Contributions by the Author

This section lists the peer-reviewed publications and open-source resources that support the thesis and promote reproducibility. The thesis is primarily based on three first-authored publications, each aligned with a core chapter. A fourth publication—a book chapter—is included as a complementary contribution. Additionally, one open-source code repository for the CAL-Tutor project and a HUP-3D project website containing the HUP-3D and HUP-3D-v2 synthetic datasets—along with multiple related repositories—are provided.

Edwards, P.; Chand, M.; **Birlo, M.**; and Stoyanov, D. (2020). *The Challenge of Augmented Reality in Surgery*. In Atallah, S. (Ed.), *Digital Surgery*, pp. 121–135.

- Cited as [52]; supporting, non-core contribution.

- Offers a critique of surgical AR research, incorporating insights from the systematic review presented in Chapter 2.

Birlo, M.; Edwards, P. J. E.; Clarkson, M.; and Stoyanov, D. (2022). *Utility of optical see-through head mounted displays in augmented reality-assisted surgery: A systematic review*. *Medical Image Analysis, 77*, 102361.

- Cited as [53]; basis for Chapter 2.
- Systematic review of OST-HMD use in surgical AR from 2013–2020, with a focus on human factors and limitations in clinical adoption.

Birlo, M.; Edwards, P. J. E.; Yoo, S.; Dromey, B.; Vasconcelos, F.; Clarkson, M. J.; and Stoyanov, D. (2023). *CAL-Tutor: A HoloLens 2 Application for Training in Obstetric Sonography and User Motion Data Recording*. *J. Imaging, 9*(1), 6.

- Cited as [54]; basis for Chapter 3.
- Describes a novel MR system for ultrasound training, integrating HoloLens 2 and a physical phantom, with recorded user motion data for future AI applications.

GitHub Repository: `https://github.com/manuelbirlo/CAL-Tutor`

Birlo, M.; Caramalau, R.; Edwards, P. J. E.; Dromey, B.; Clarkson, M. J.; and Stoyanov, D. (2024). *HUP-3D: A 3D Multi-View Synthetic Dataset for Assisted-Egocentric Hand–Ultrasound Probe Pose Estimation*. In *Proceedings of MICCAI 2024*, LNCS 15001, Springer, pp. 430–436.

- Cited as [55]; basis for Chapter 4.
- Presents a synthetic RGB-D dataset with egocentric/non-egocentric views for hand and probe pose estimation, including evaluation with state-of-the-art hand and tool pose estimation DL model *HOPE-Net*.

Dataset and Code: `https://manuelbirlo.github.io/HUP-3D/`

Summary

This introductory chapter established the framework for the thesis by introducing medical AR, outlining its motivation, historical evolution, and key challenges—particularly for OST-HMDs. It then stated the aims, main contributions, and thesis structure, and listed the related publications. In brief, the thesis contributes: (i) a systematic review of OST-HMD use in surgery, highlighting technical and human-factor barriers; (ii) CAL-Tutor, a HoloLens 2-based MR training prototype for obstetric US, with a user study demonstrating feasibility and highlighting limitations of marker-based probe tracking; and (iii) HUP-3D/HUP-3D-v2, synthetic multi-view datasets (RGB, depth, and segmentation maps) with evaluations for markerless hand-probe pose estimation. With this foundation, Chapter 2 presents the systematic review.

Chapter 2

Assessing the Utility of OST-HMDs in AR-Assisted Surgery: A Systematic Review

2.1 Introduction

This chapter presents a systematic review of the current state of AR in surgical applications, with particular emphasis on the use of OST-HMDs. The goal of this review is to assess the landscape of AR-assisted surgery, identify common limitations—both technical and human factor-related—and highlight key trends and research gaps in the field. Most of the content in this chapter is based on the author's peer-reviewed publication in Medical Image Analysis [53], with minor adaptations and reformatting for consistency within the thesis.

The chapter begins by establishing the motivation for conducting this review, followed by a detailed description of the review methodology, inclusion criteria, and analysis strategy. The results are then categorised and discussed across dimensions such as device type, surgical specialty, application context, and evaluation approach. Chapter 1 provided a broad overview of the evolution and current challenges in medical AR, with a particular focus on the growing research interest in OST-HMD-based surgical systems. Despite an increasing number of publications, the clinical adoption of these systems remains limited due to persistent technical and human factor challenges. To contextualize the relevance of this systematic review, it is useful to examine how the release of commercial OST-HMDs has influenced research activity in surgical AR. Devices such as Google Glass and Microsoft HoloLens 1 and 2 have played a key role in driving a significant increase in related publications over the past decade. Figure 2.1 illustrates the distribution of surgical AR articles involving OST-HMDs published over the last 24 years.

As outlined earlier, this chapter builds on a systematic literature review conducted by the author and co-authors [53], which covered publications up to the end of 2020. Since the release of the first HoloLens in 2016, the device has become a dominant platform in surgical AR research—an upward trend that continued with the launch of HoloLens 2 in 2021 [41]. More recently, the introduction of the Apple Vision Pro in 2023—representing a VST-HMD rather than an OST-HMD—has drawn attention as a potential alternative for surgical applications [56, 57, 58, 59]. However, given their distinct underlying technology and interaction paradigms, VST-HMDs are beyond the scope of this review. The remainder of this chapter outlines the evolution and clinical application of OST-HMDs in surgery, followed by a detailed analysis of core limitations such as spatial registration accuracy, perceptual 3D virtual object accuracy, and human factors of human computer interaction. The goal is to identify unresolved challenges and research gaps that serve as a foundation for the methodological and experimental contributions presented in Chapters 3 to 5.

As OST-HMDs have driven much of the recent growth in surgical AR research, it is important to understand how they compare to other types of HMDs, particularly VST-HMDs. AR systems can be categorised by their underlying display technology, which includes conventional monitors, projectors, and HMDs [24, 60]. Among these, HMDs are generally considered the most user-friendly for manual tasks, as they allow for hands-free interaction from a self-centered perspective[42]. HMDs can be further classified into two main categories based on their augmentation method: VST-HMDs and OST-HMDs.

In Video See-Through (VST) systems, a video image feed is combined with

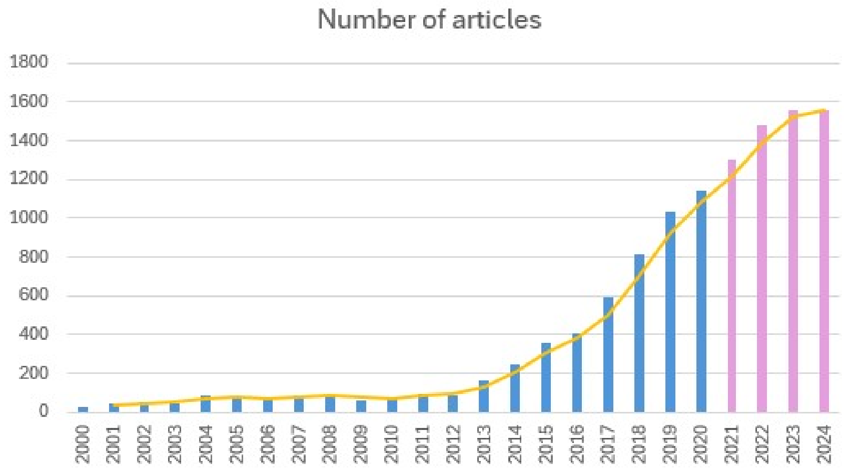


Figure 2.1: Distribution of number of articles within the last 24 years that address surgical AR HMD-related topics: Google Scholar search results for surgery ”Head Mounted Display” ”Augmented Reality” OR ”Mixed Reality” surgery ”Head Mounted Display” ”Augmented Reality” OR ”Mixed Reality” ”optical see through” OR ”Hololens” OR ”Magic Leap” OR ”Google Glass” in the last 20 years. Search performed on 24 March 2025. In the literature review paper published by Birlo et al. [53], only publications up to 2020 were included. The distribution of number of articles from 2021 to 2023 confirms the continuation of a still ongoing upwards trend, with a plateau in 2024.

superimposed computer generated images, such as 3D reconstructed MRI scans of organs. VST systems have been adopted in surgical applications via computer displays and HMDs, and offer potential advantages such as improved synchronisation between video feed and overlay and video processing for image segmentation or registration. In addition, the contrast between video feed and virtual overlay can be easily controlled, allowing the virtual overlay to occlude the real scene, which is not the case for optical see-through (OST) systems. On the other hand, VST systems face disadvantages including limitations in terms of video bandwidth, the risk of losing vision of the real scene in the case of system errors, and geometric aberrations such as distorted spatial perception [61]. Though video and overlay may be well synchronised, there is inevitably some delay between actual motion and perception of the motion, both real and overlaid, which can slow down surgical motion and may increase errors. [62] also noted that the absence of a direct view to the real world makes surgeons nervous.

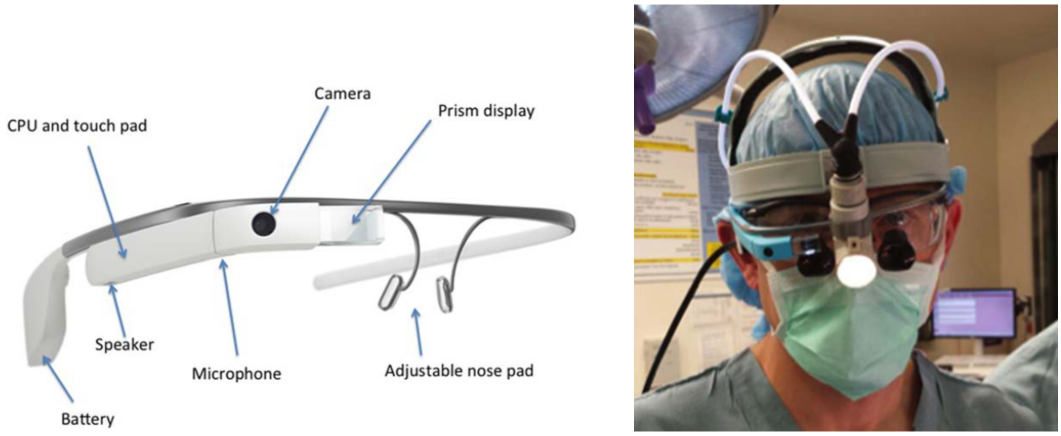
In OST-HMDs a transparent monitor displaying 3D-rendered content is located between the surgeon's line of vision and the target anatomy. This ensures an unhindered view of reality, i.e. natural stereo vision capabilities without lag or loss of resolution associated with the real surgical scene. However, the drawbacks include dynamic registration errors for the augmented view, latency when moving static registration errors, complex calibration and unnatural perceptual issues. For example, nearer virtual objects don't occlude real objects in the background [63].

To contextualise the capabilities and limitations of currently used OST-HMDs, it is helpful to consider their hardware evolution—from custom-build prototypes tailored for surgical use cases to modern commercial devices originally design for general MR applications. Commercial HMDs dominate recent AR research due to their ergonomic design and hands-free, user-centered interaction, which are essential for surgical tasks. Nevertheless, their origins outside of surgical use create unique challenges—including registration accuracy, occlusion handling, and sterility considerations—that must be carefully evaluated. The following section provides a brief background on key developments in OST-HMD hardware that have shaped surgical AR, which dedicated subsections addressing both safety and clinical efficacy, two crucial aspects of their adoption in medical practice.

2.2 Background on OST-HMDs in Surgical AR

Before 2013, OST-HMD-based research largely relied on custom build devices. Creating such a custom device is a technically difficult challenge, miniaturised displays into a wearable headset with half-silvered mirrors enabling free view of the real scene. It is hard to achieve an optical setup to display a bright image with good contrast and resolution covering a wide field-of-view. Therefore, it is not surprising that only two of the papers in this review use such custom devices.

Commercially driven benefits through the ability to place graphical information directly overlaid on the wearer's view of the real world has led to the development of a number of commercial devices. Google Glass, released in 2013, is a lightweight monocular AR device enabling display of information while allowing



(a) **(b)**

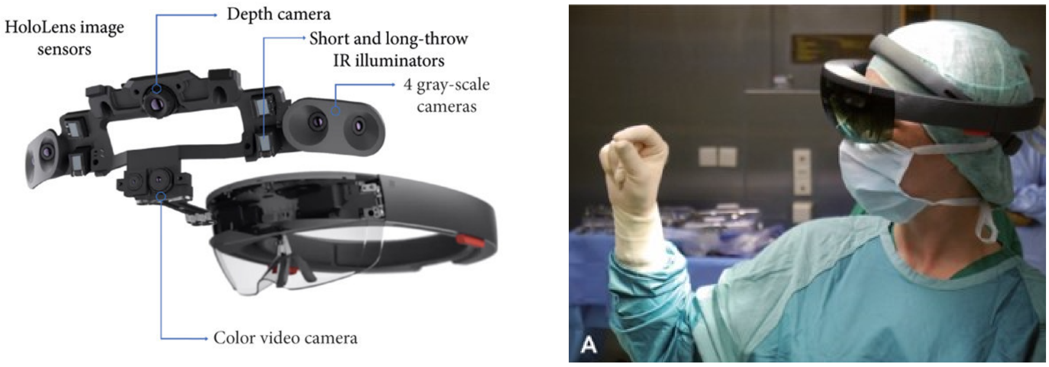
Figure 2.2: Google Glass in a surgical education setup, used for intratoperative consultation via telecommunication in otolaryngology surgery. (a) Overview of the Google Glass device, illustrating its components, including the camera, microphone, speaker, and prism display. Figure reused from Moshtaghi et al. [64]. (b) Surgeon wearing the Google Glass along with separate magnification loupes and a headlight. Figure adapted from Moshtaghi et al. [64].

users to continue daily tasks without needing to look away at an external screen or use their hands. Fig. 2.2 shows the Google Glass in surgical education setup. The Microsoft HoloLens, released in 2017, offers most of the benefits of the Google Glass and is a larger HMD that incorporates stereo vision, low latency room mapping and head tracking as well as gesture-based interaction using only the wearer's hands. Fig. 2.3 illustrates the hardware components of the HoloLens and its use in an intraoperative setup.

Numerous other devices have appeared offering different levels of comfort and function (for a more detailed list of OST-HMD devices see section 2.5.3). Although none of these devices were specifically designed for surgical tasks, the potential for a convenient display of information to the surgeon has led to a significant increase in research detailed in this review. In common with any medical intervention, the fundamental questions concern safety and efficacy. These are discussed in Subsections 2.2.1 and 2.2.2.

2.2.1 Safety of OST-HMDs in Surgical AR

Safety is one of the most important considerations when evaluating the use of OST-HMDs in surgery, as device-related or human–computer interaction issues can ul-



(a) (b)

Figure 2.3: (a) Hardware components of the Microsoft HoloLens, including sensors, cameras, and display system. Figure reused from Galati et al. [65]. (b) Application of the HoloLens during visceral surgery, operated via hand gestures. Figure adapted from Sauer et al. [66].

timately affect surgical outcomes and pose risks to patient safety. Perceptual accuracy, human-factor considerations in human-computer-interaction (ergonomics etc.) and system reliability are all important factors. In particular, inaccurate spatial registration—discussed in detail in Section 2.5.6—can mislead visual interpretation of 3D virtual content and compromise manual surgical precision. This subsection briefly discusses two of the most critical safety-related challenges: perceptual accuracy and human factors.

Perceptual accuracy. The convenient overlay of surgical assistive information entails certain risks. Where the aim is that the overlay directly guides surgery, perceptual accuracy is key. Some authors are critical of OST-HMD device accuracy. [42] concluded from their quantitative study that the HoloLens should not be used for high-precision manual tasks. [67] also conclude that OST-HMDs are unsuitable for surgical guidance, suggesting that research should focus on addressing perceptual issues that play a critical role in limiting user accuracy. [68] performed a systematic review of augmented reality in open surgery and concluded that such perceptual issues limit their usage to the augmentation of simple virtual elements such as models, icons or text.

Human-factor challenges. Even precise overlay could distract from or hamper the surgeon's view of the patient, potentially slowing the response to critical situations such as bleeding. As a possible solution, [69] propose nearby presentation of correctly oriented but not registered models. Besides visual distraction, gesture interactions with the AR view may prove difficult to combine with the manual surgical task itself [70]. Cognitive overload can occur if too much extra information is presented to the surgeon at the same time [71].

[72] analysed the effects of MR HMDs on cognitive and physiological functions during intellectual and manual tasks that last for 90 minutes. Their experiment consisted of 12 volunteers performing and manual tasks with and without the HoloLens while their physical and mental conditions (cognitive, cardiovascular and neuromuscular) were measured. They conclude that using the HoloLens is safe since it does not impact safety-critical human functionalities like balance and cognitive and physical fatigue. However, despite the positive outcome of the study, the authors also state that one of the prerequisites of a safe and effective usage of HMDs is that users should be receptive to the device.

Addressed and unresolved challenges. While some hardware limitations of OST-HMDs, such as narrow field of view and calibration accuracy, have been partially addressed in newer models like the HoloLens 2 (Microsoft Corporation, Redmond, USA), the aforementioned perceptual accuracy and human-factor challenges remain a major barrier to clinical adoption. These issues—covering perception, cognitive load, and device acceptance—continue to limit the commercial success of OST-HMD-based AR in surgical applications [61, 67]. When considering the adoption of OST-HMD-assisted methods into surgical routine, efficacy is another key consideration, which is addressed in the next subsection.

2.2.2 Efficacy of OST-HMDs in Surgical AR

While safety remains a foundational requirement for OST-HMDs deployment in surgery, efficacy is equally critical to justify clinical integration. These devices offer the benefit of displaying graphical elements—such as images, icons, or text—directly within the surgeon's field of view. There is no need to look away from the surgical scene or stop the operation to obtain potentially useful visual input. However, it is important to be aware of potential system-related drawbacks, especially during high-precision surgical tasks such as surgical guidance. When displaying guidance information, accuracy becomes a measure of system performance and the majority of the papers included in this review perform some accuracy or precision experiments. As detailed in the previous Subsection 2.2.1, it is important to distinguish registration or tracking accuracy, which is often based on an external tracking or guidance system, from perceptual accuracy achieved by the AR system. Human factor limitations might be equally important since they directly or indirectly impact system performance. For example, a higher cognitive load might lead to a decreasing task performance of increased risk of errors. Likewise, discomfort when wearing an OST-HMD can cause unwanted physical long term effects that in turn may impact the user's ability to focus on the task at hand.

In general, efficacy of a proposed OST-HMD solution is measured via a variety of metrics. Such evaluation metrics comprise quantitative metrics from studies, e.g., error rates or time reduction, usability assessments that bridge safety and efficacy, study designs (simulations, cadaver studies, clinical trials, etc.) and OST-HMD device comparisons (HoloLens 1 vs. 2, etc.). Appendix table A4—which lists the used AR visualisation, conducted experiments and reported accuracy of all final articles included in the meta-analysis of this literature review—includes some of these evaluation metrics. One could argue that the ultimate test of efficacy would be improved patient outcome, but the systems reviewed are not currently at the stage of large-scale clinical trials that would be needed to demonstrate patient benefit. In summary, both safety and efficacy considerations highlight the complexity of integrating OST-HMDs into surgical workflows and the importance of critically assessing the current body of research in this domain.

2.3 Methods

To address the challenges and gaps identified in the previous section, the following outlines the methods used to conduct a systematic literature review on OST-HMD-assisted surgical applications. Section 2.3.1 describes the literature search strategy used to retrieve the initial set of articles screened for further relevance. To contextualise this review, additional thematic literature review articles are briefly introduced in Subsection 2.3.1.1. Section 2.4 then details the Google Scholar analysis strategy applied to identify relevant studies for inclusion in the meta-analysis. A structured review protocol was followed, incorporating clearly defined inclusion and exclusion criteria, and focusing on peer-reviewed publications related to OST-HMD applications in surgery. Following the systematic review process, a total of 91 articles were identified as relevant and analysed in Section 2.5. These articles were categorised according to various aspects, including surgical speciality, OST-HMD device type, application context, and experimental validation. The section concludes with a detailed discussion of human factors, highlighting their complementary role alongside the technological aspects of proposed AR-assisted surgical systems. Additional tables providing detailed information on specific categorisations of the 91 included articles are available in Appendix Chapter A.

2.3.1 Literature Search and Scope

This systematic review, originally published in [53], provides an overview of recent advancements in OST-HMD-assisted surgery, outlining major contributions and innovations up to 2020. Consequently, publications between 2021 and 2024 are not part of the meta-analysis of the literature analysis strategy (Section 2.4 and literature search 2.5). The analysis is conducted by inclusion of several components of the selected literature, including OST-HMD device, surgical speciality, surgical application context, surgical procedure, AR visualisations, conducted experiments and accuracy results. A special focus is given to the identification of human factors in each article.

A systematic review was performed according to the preferred reporting items for systematic review and meta-analysis (PRISMA) guidelines [73]. The literature search was conducted on Google Scholar with the search terms [surgery "Head Mounted Display" "Augmented Reality" OR "Mixed Reality" surgery "Head Mounted Display" "Augmented Reality" OR "Mixed Reality" "optical see through" OR "Hololens" OR "Magic Leap" OR "Google Glass"]. An initial Google Scholar including all articles between 2013 and 2020 was conducted on February 21, 2020.

An updated Google Scholar search for 2020 only was subsequently performed on January 27, 2021.

This review covers only original research papers; other literature review papers are not considered. The original literature search (section 2.4) did return a number of these, however, which deserve some attention, a few of which are mentioned in the following Subsection 2.3.1.1. This section also covers a few more recent relevant review papers that were not discovered during the original literature search due to their more recent publication dates (2021-2024), but are worth mentioning and are discussed under *Complementary literature* of Subsubsection 2.3.1.1. The results of the literature search are presented in Section 2.5.

2.3.1.1 Contextual Review Literature

Although not part of the formal meta-analysis, other relevant literature review papers exist that deserve to be mentioned and reveal recurring barriers to the adoption of AR-assisted surgical methods. A general review of all areas of AR, including medical and surgical, is provided by [74], who examine the usability of AR over a 10 year period. [75] review medical applications of MR and provide a broad taxonomy. A comprehensive review of medical AR is provided by [76] who conclude that there is no proof of clinical effectiveness as yet. [77] give a comprehensive review using the Data–View–Visualization (DVV) taxonomy and provide suggestions for areas that need attention, including specific overlays for important phases of the operation as well as optimisation of interaction and system validation. A previous publication, co-authored by the author of this thesis, identified several barriers to the adoption of surgical AR [52]. Existing comprehensive reviews of related surgical areas were found, including robotics [78] and laparoscopic surgery [79]. Orthopaedics is the dominant application area in this review and three other reviews cover this specific field well [80, 81, 82].

Complementary literature. Several newer literature reviews (2021-2024) expand on these general findings. [83] provided a review on extended reality (XR)-based HMDs, including AR, MR, and VR, and noted a general benefit for medical education with a focus on teaching skills and knowledge. However, the review also criticized the lower range of diverse educational settings that were analysed by researchers due to a concentration on small-scale studies in high-income countries, suggesting a research expansion to low- and middle-income countries. [47] reviewed challenges of OST-HMDs in AR-assisted surgery and found that while there are promising tendencies in terms of navigation accuracy and usability, technical and human factor challenges, such as perception, ease of use, and interaction still need to be further improved before widespread clinical acceptance and routine practice can be achieved. [84] investigated trends of HMD-based VR and AR systems in medical education and found that while these technologies show a significant potential to enhance clinical training by increasing student motivation and satisfaction, further research is needed to establish standardised protocols and a more diverse effectiveness validation. [85] focused their review on VR and AR in spine surgery and concluded that while AR and VR show the potential to improve surgical training, planning and intraoperative guidance, further research is needed to effectively confirm clinical utility and overcome technical and ergonomic hurdles. [56] compared the performance of using OST-HMDs instead of conventional monitors during laparoscopic surgery in an experimental setting. They found that the main obstacle preventing OST-HMDs from replacing conventional monitors is the device's weight, which causes physical discomfort when worn for extended periods. However, they also showed that under certain experimental conditions, surgical task performance and situation awareness were better than in the conventional monitor setting. [86] reviewed the state-of-the-art of OST-HMD applications in context-aware systems for open surgery. They found that the AR display of automatically analysed, context-aware information improves users' task completion and decision-making capabilities. However, they also note that current OST-HMDs have limitations, such as a restricted field of view, discomfort due to the device's weight, and an increased likelihood of disrupting the surgical workflow due to the need for frequent re-calibration to maintain accurate alignment of 3D virtual content.

While these review papers were not included in the formal meta-analysis presented in Section 2.4, they offer consistent insights. Across both earlier and more

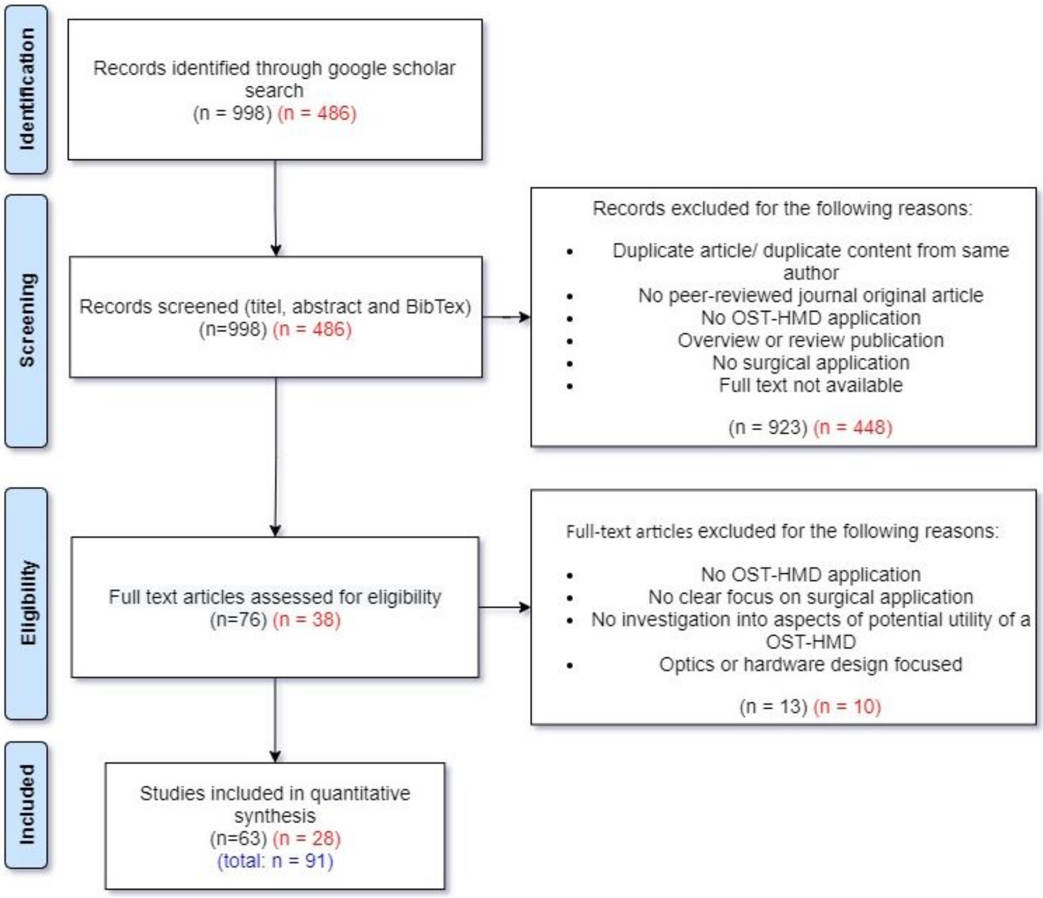

Identification
Screening
Eligibility
Included
Records identified through google scholar search
(n = 998) (n = 486)
Records screened (titel, abstract and BibTex)
(n=998) (n = 486)
Records excluded for the following reasons:
Duplicate article/ duplicate content from same author
No peer-reviewed journal original article
No OST-HMD application
Overview or review publication
No surgical application
Full text not available
(n = 923) (n = 448)
Full text articles assessed for eligibility
(n=76) (n = 38)
Full-text articles excluded for the following reasons:
No OST-HMD application
No clear focus on surgical application
No investigation into aspects of potential utility of a OST-HMD
Optics or hardware design focused
(n = 13) (n = 10)
Studies included in quantitative synthesis
(n=63) (n = 28)
(total: n = 91)

Figure 2.4: Systematic review search strategy

recent reviews, there is broad agreement on the potential of AR in surgical applications, alongside recurring concerns about technological limitations, user acceptability, and the lack of clinical validation. Notably, although a few recent reviews have begun to address OST-HMDs, comprehensive analyses focusing exclusively on these devices remain limited—particularly in relation to human factors and clinical adoption.

2.4 Literature Analysis Strategy

To narrow the article search scope by publication year and focus on recent research, the number of publications resulting from Google Scholar search terms over the last 20 years at the time of writing the literate review paper was analysed (Fig. 2.1, Chapter 1). This shows a steady increase starting in 2013, coinciding with the release of Google Glass. Due to this noticeable increase in relevant publications,

Table 2.1: Inclusion criteria used during abstract and full-text screening phases, and data extracted from included publications

Abstract Screening Criteria
1. Peer-reviewed original journal article
2. OST-HMD focused application with surgical context
3. Not an overview or systematic review publication
Full-Text Screening Criteria
1. Describes the usage of an OST-HMD
2. Clear focus on a surgical application
3. Investigates the potential utility of OST-HMDs in surgical settings
4. Not focused on optics or hardware design
Data Extracted from Included Publications
1. Clinical setting (surgical specialty, surgical application context, surgical procedure)
2. The assessed OST-HMD device
3. Methods (AR visualisations, conducted experiments)
4. Key results (accuracy)
5. Human factors

the year 2013 was selected as the starting point for this literature review.

The review process is shown in Fig. 2.4 and includes the results from both the original search (February 21, 2020, numbers in black colour) and the updated search (January 21, 2021, numbers in red colour). The Google Scholar search initially resulted in 998 (486) records. In a subsequent screening phase, title, abstract and BibTex information were read to decide whether the record seems to be a relevant publication. Records identified as duplicates, or containing substantially duplicated content from the same authors, were excluded. A total of 15 (7) duplicates were excluded. During the screening, the three inclusion criteria listed under **Abstract Screening Criteria** in Table 2.1 were applied.

Records whose full text wasn't available were excluded. 907 (441) records that didn't meet the inclusion criteria were excluded. Together with the 15 (7) excluded duplicates, a total of 923 (448) records were excluded during the screening phase, which led to 76 (38) remaining full text articles that were assessed for eligibility. Full text articles had to meet the four inclusion criteria: The article listed under

Full-Text Screening Criteria in Table 2.1. 13 (10) articles that didn't meet these inclusion criteria were excluded. The remaining 63 + 28 = 91 studies that met all predefined inclusion criteria form the final set of papers examined in this review.

When reporting the results, the PRISMA guidelines were followed. Due to the inherent characteristics of the studies (small case series, subjective qualitative assessments, no controlled randomised trials) a meta-analysis could not be performed. Therefore, publication bias could not be reduced and should be taken into account. Data extracted from the included publications are listed under **Data Extracted from Included Publications** in Table 2.1 and detailed in the following Section 2.5.

2.5 Analysis of the Literature Search

This section summarises the results of the included 91 articles identified through the systematic literature review. The analysis is structured into several subsections, each focusing on a specific aspect of the reviewed articles to provide a comprehensive overview. The following subsections analyse: (i) the annual distribution of used OST-HMD devices (Subsection 2.5.3); (ii) the distribution of surgical specialities (Subsection 2.5.2); (iii) the distribution of surgical application contexts (Subsection 2.5.4); (iv) types of AR visualisations used (Subsection 2.5.4.3); (v) experimental evaluations of AR (Subsection 2.5.5); (vi) registration and tracking in surgical AR (subsection 2.5.6); and (vii) human factors (Subsection 2.5.7).

A special emphasis is placed on human factors, which contributes to the novelty of this literature review. Several human factors were identified and grouped into three sub categories: information perception, cognitive processing and control action. Additional analysis details regarding specific aspects of the reviewed literature can be found in Appendix Chapter A: An overview of the OST-HMD devices used, surgical application contexts, and surgical procedure is provided in appendix Table A2. Appendix Table A4 contains details about AR visualisations, conducted experiments and accuracy results.

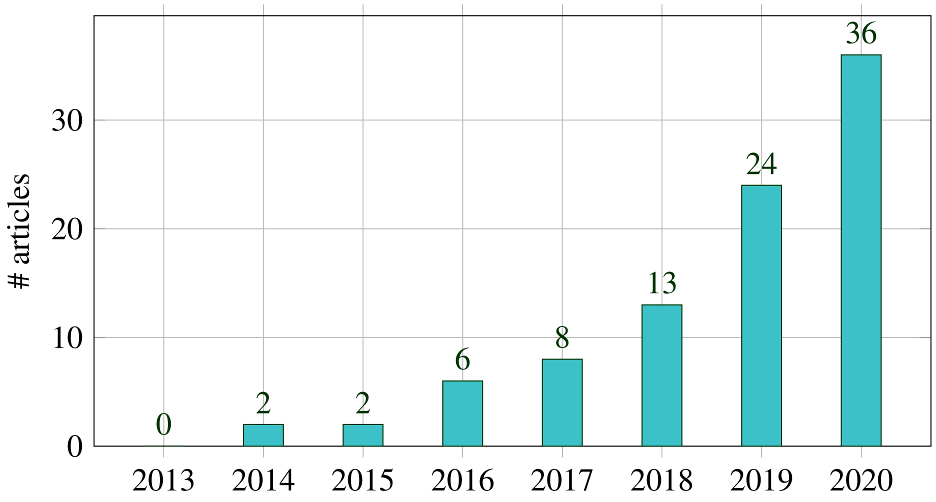


Figure 2.5: Systematic review results overview: Annual Distribution of selected 91 studies from 2013-2020

2.5.1 Annual Distribution of Selected Articles

The first aspect examined was the temporal distribution of the articles that met the inclusion criteria for this literature review. This trend corresponds with the broader increase in publications shown in Fig. 2.1 (Chapter 1), which reflects general research activity on surgical HMDs. However, the distribution presented here (Fig. 2.5) is restricted to the 91 articles included in the systematic review and thus represents a refined selection of relevant literature. As shown, there were no included articles in 2013; beginning in 2014, the number of relevant publications rose and generally continued to increase, culminating in the highest annual count in 2020. This trend likely correlates with the commercial release of major OST-HMDs such as Google Glass and Microsoft HoloLens, indicating that hardware availability plays a key role in enabling OST-HMDs-related surgical research.

The trend observed in the selected 91 articles closely mirrors the overall increase in published articles related to surgical OST-HMDs, as shown in Figure 2.1 (Section 2.1). Notably, the broader data presented in Figure 2.1 shows a rise in publication from 2013 onwards, with a plateau in 2024. This alignment supports the representativeness of the selected articles within the context of the wider research landscape. Furthermore, despite the literature review by Birlo et al. [53] only covering publications up to 2020, the data in Figure 2.1 confirms a continuing upward trend through 2023. This suggests a sustained and growing research interest in the field. It is therefore reasonable to assume that the number of identified papers would have continued to increase beyond 2020 if the literature review had included papers between 2021 and 2024. The following subsections analyse how the 91 selected articles are distributed across different categories, such as surgical specialty, device type, and application context.

2.5.2 Surgical Speciality

This review found that OST-HMDs have been applied in a variety of surgical specialities. Fig. 2.6 shows a graphical illustration of all articles grouped into their surgical speciality and placed on the respective body region, depicted on a schematic image of the human body. Fig. 2.7 shows the proportion of publications for each surgical speciality. Orthopaedic surgery dominates (28.6%, n = 26), perhaps since proximity to bone requires only rigid registration and somewhat lower accuracy is required compared to applications such as neurosurgery. General surgery, neurosurgery, applications without a concrete surgical speciality and vascular surgery follow with more than five articles each. Dental surgery is represented with five articles, followed by heart surgery and Otolaryngology (n = 4 each). Other surgical specialities include reconstructive surgery, urology and maxillofacial surgery (n = 3 each). A few attempts have been made to explore potential benefits of OST-HMDs in robot-assisted surgery and paediatric surgery (n = 2 each). Interventional oncology, laparoscopic surgery, visceral surgery and anaesthesiology are represented with one article. Specific articles per surgical speciality are detailed in table 2.2. While orthopaedics still dominates, other applications, including general, vascular and neurosurgery, are increasingly represented in the latter half of the survey period as interest in AR applications spreads to other surgical fields.

2.5.3 Device Type

Fig. 2.8 depicts the annual distribution by OST-HMD between 2014-2000. Google Glass, the device with the second highest number of articles (n = 8), dominates the distribution in 2014, but interest decreases from 2015 to 2017, perhaps due to

Table 2.2: Distribution of the included articles per surgical speciality

Surgical speciality	Articles
Orthopaedic Surgery	[87] [88] [89] [90] [91] [92] [93] [94] [95] [96] [97] [98] [99] [100] [101] [102] [103] [104] [105] [106] [107] [108] [109] [110] [111] [112]
General Surgery	[113] [114] [115] [116] [117] [118] [119] [120] [81] [65] [121]
Neurosurgery	[122] [123] [124] [125] [126] [127] [128] [129] [130]
Vascular Surgery	[131] [132] [133] [134] [135] [136] [137]
General surgical applications	[138] [139] [140] [62] [141] [142] [143] [144]
Dental Surgery	[71] [145] [146] [147] [148]
Heart Surgery	[149] [150] [151] [152]
Otolaryngology - head and neck surgery	[153] [154] [155] [156]
Reconstructive Surgery	[157] [158] [159]
Maxillofacial Surgery	[160] [161] [162]
Urological surgery	[163] [164] [165]
Robot-assisted surgery	[166] [167]
Paediatric surgery	[168] [169]
Visceral Surgery	[66]
Interventional Oncology	[170]
Laparoscopic Surgery	[171]
Anaesthesiology	[172]

diminishing support from Google. The Microsoft HoloLens was released in 2016 and has dominated the field of OST-HMD assisted surgery since then, with a steady increase in papers from 2017 and accounting for the majority of articles (n = 66). Following HoloLens and Google Glass, the Moverio BT-200, which was released in 2014, has the third highest number of articles (n = 4) and was used once in 2016, 2017, 2019 and 2020. Its successor, the Moverio BT-300, was released in late 2016 and has only one application in 2020. The Magic Leap One, released in 2018 and attracting huge initial investment, has not established itself in OST-HMD assisted surgery, generating only one article in 2019. Other devices include the NVIS nVisor ST (n = 2), Vuzix M300, Brother AirScouter WD-100, Aryzon headset, Metavision Meta 2 and PicoLinker (n = 1 each).

To summarise, by the time this literature review was conducted, the HoloLens clearly dominated the field, although there was growing interest in other devices such as the Moverio BT. The field was still rapidly evolving, driven in part by the release of major OST-HMDs, including the HoloLens 2 in 2019. HoloLens 2-related surgical applications began gaining popularity in 2021. In 2020, only a few HoloLens 2-related articles were published, primarily within medical but not directly surgical contexts [41]. In 2023 the Apple Vision Pro (Apple Inc., Cupertino, CA, USA) was released, a VST-HMD that is gaining attention in the research

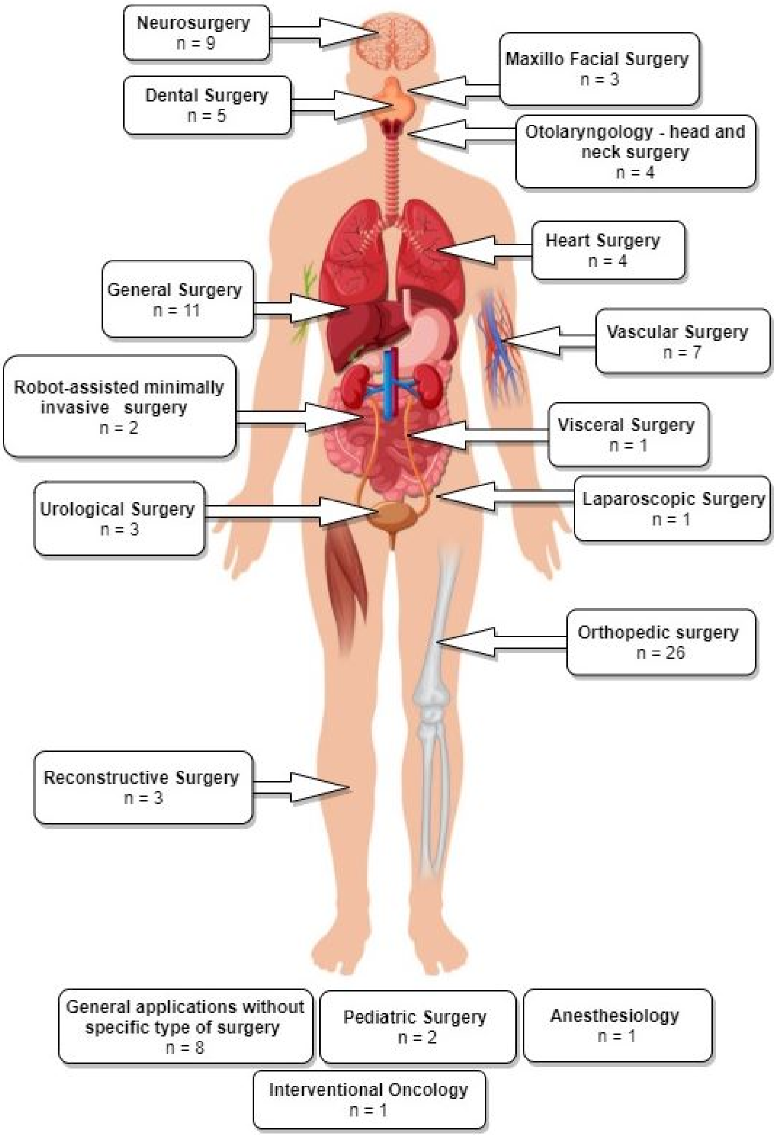


Figure 2.6: Graphical illustration of included articles grouped by surgical speciality and placed at respective human body regions

community [173, 57, 59, 58]. However, since it is a VST-HMD and not a OST-HMD, it will not be considered in this chapter.

2.5.4 Surgical Application Context

Surgical application contexts define how OST-HMD assistance is intended to improve surgical practice. Fig. 2.9 shows the distribution of all identified contexts. Surgical guidance is by far the most popular (n=58), followed by preoperative surgical planning (n=12) and surgical training (n = 11) then Teleconsultation and tele-

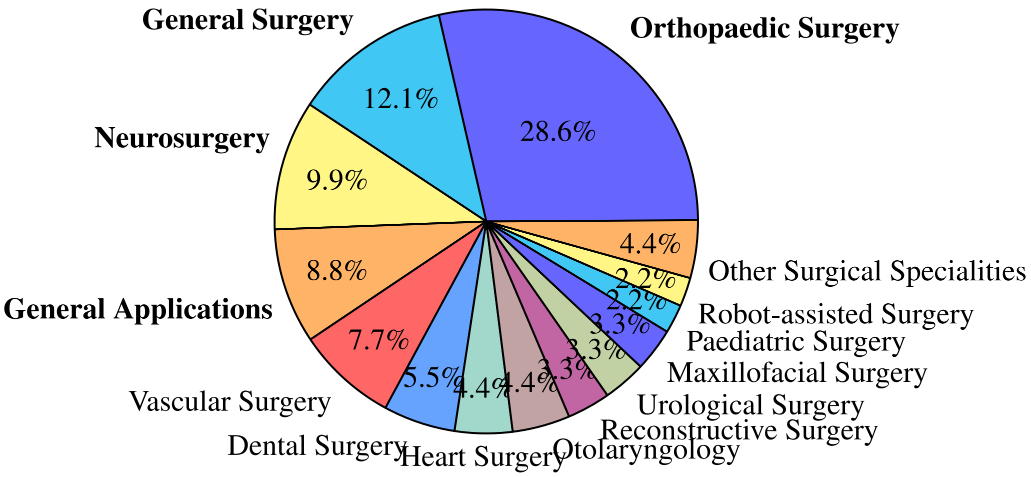


Figure 2.7: Pie chart showing the distribution of the included 91 papers across surgical specialities

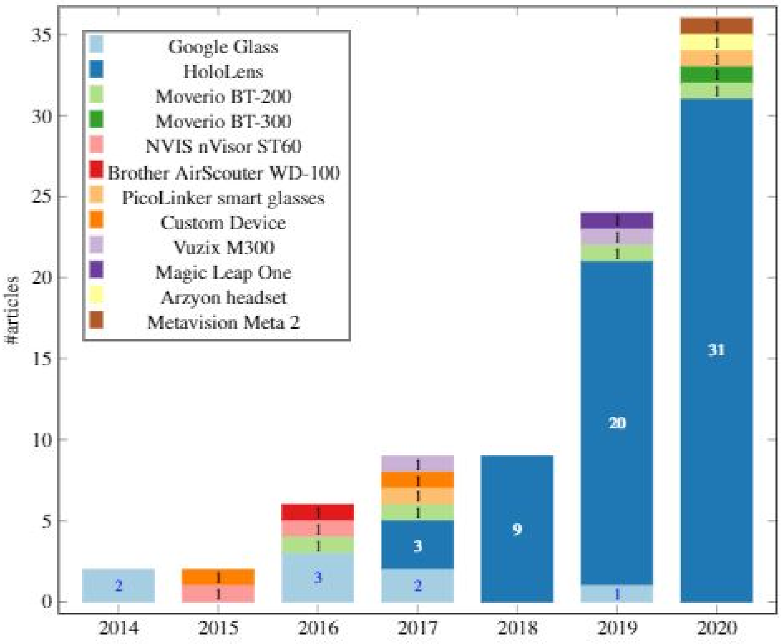


Figure 2.8: Annual Distribution of articles by OST-HMD device from 2014-2020

mentoring (n = 5 each). Four articles were included in which the surgeon views a 3D patient anatomy 3D virtual object to aid clinical decision making, rather than for intraoperative guidance. This category is referred to as *intraoperative surgical anatomy assessment*.

The remaining applications that have been identified are intraoperative review of preoperative 2D imaging and/or patient records, intraoperative documentation,

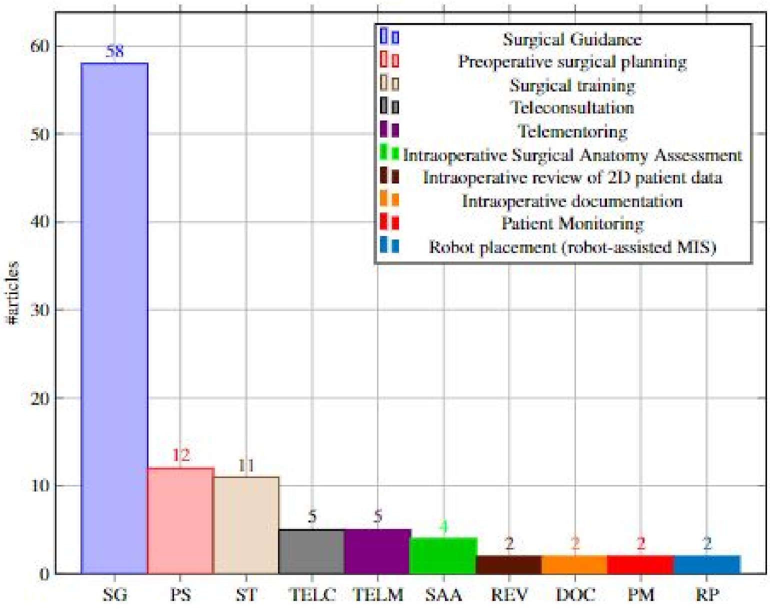


Figure 2.9: Distribution of articles by **surgical application context**

patient monitoring and preparation of robot-assisted MIS (n = 2 each). Several of the surgical application contexts and corresponding articles are further elaborated in the following subsubsections. Surgical guidance is the most popular surgical application context for exploring OST-HMD-assisted methods, and its distribution accross the selected literature is therefore described in a dedicated subsubsection. The distribution of the remaining surgical application contexts—including preoperative surgical planning, surgical training, and teleconsultation—is presented in the subsequent subsubsection.

2.5.4.1 Surgical Guidance

Surgical guidance, also known as image-guided surgery, is defined by Cleary and Peters [174] as a medical procedure in which a surgeon uses computer-based virtual pre- or intraoperative image overlays to visualise and target patient anatomy. The authors also state that an image-guided intervention includes registration and tracking methods. However, an OST-HMD-based solution can also be considered a form of image guidance if it employs registered 3D virtual image overlays—regardless of tracking—as long as these overlays assist the clinician in visualising and targeting the surgical site. Since this broad definition encompasses over half of the included

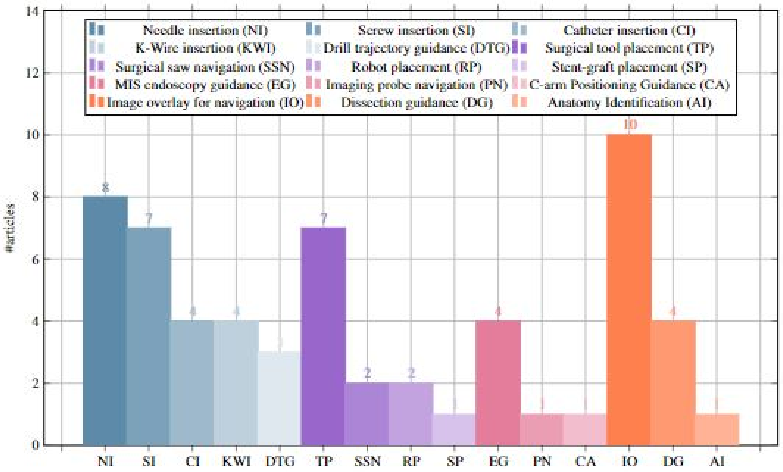
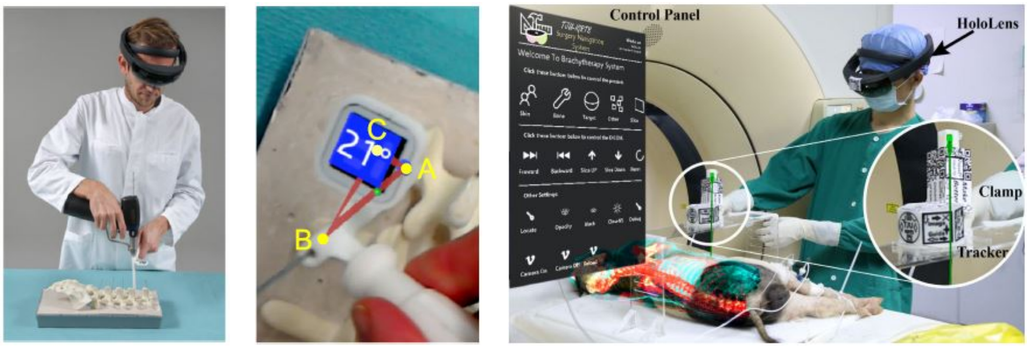


Figure 2.10: Surgical guidance applications: Distribution of the subset of final 91 articles (n = 59) by applications of **surgical guidance**, grouped into the four categories 1. navigation of a linear path, 2. navigation of surgical tools or equipment, 3. navigation of an imaging device, 4. general guidance to help spatial awareness not associated with a specific task

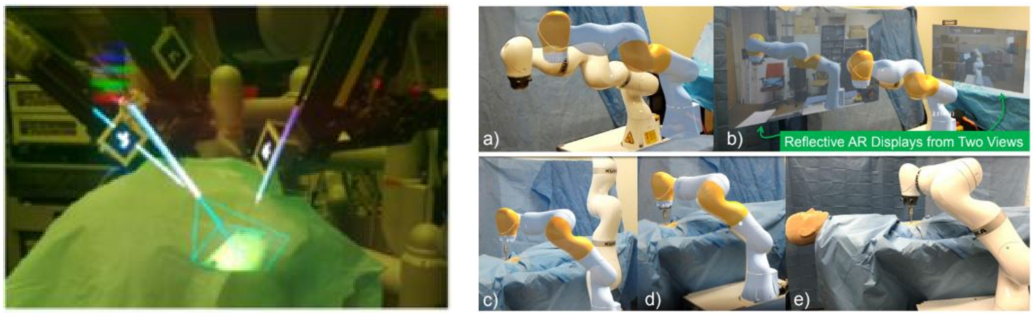


(a) (b) (c)

Figure 2.11: Guided screw insertion and needle insertion examples. (a) A surgeon uses a custom-made navigation device in an experimental setup (b). Augmented drill entry points (shown in blue) are used to start the navigation. During the guided drill procedure, the 3D angle between current and targeted screw trajectory and their deviation angle are displayed. *(source: [102] Fig. 5b and 5d)*. (c) MR needle insertion navigation system for low dose-rate (LDR) brachytherapy *(source: [133]) Fig. 1.*

papers, OST-HMD-assisted surgical guidance was further divided into distinct application categories, the distribution of which is presented in Fig. 2.10.

General image overlay for navigation systems (n = 10) overlay a registered 3D anatomy model in order to provide surgical guidance, including applications in neuronavigation [124, 125], orthopaedic procedures [100], algorithm-focused registration approaches [114, 101, 139] and maxillo-facial tumor resection [161]. Needle insertion (n = 8) has emerged as an application since 2018, mostly using the HoloLens, and was investigated in percutaneous spine procedures [96], needle biopsy [113], thoracoabdominal brachytherapy [133, 119] and needle-based spinal interventions [127]. [133] presented a mixed reality based needle insertion navigation system for low-dose-rate brachytherapy that was tested in animal (Fig. 2.11 (c)) and phantom experiments. Reported benefits of this needle insertion approach include clinically acceptable needle insertion accuracy and a reduction of the number of required CT scans. Tool placement examples (n = 7) include investigated attentiveness to the surgical field during navigation [91], a first assistant's task performance during robot-assisted laparoscopic surgery based tool manipulation [166], bone localisation [95], an optical navigation concept [138], liver tumor puncture [117], craniotomy assistance [126] and percutaneous orthopaedic treatments [104]. OST-HMD assisted screw insertion (n = 7) has been explored with different 3D virtual visualisations. [122] presented an application for pedicle screw placement in spine instrumentation that streamed 2D neuronavigation images onto a Google Glass. Surgeons reported an overall positive AR-experience. [102] developed a HoloLens pedicle screw placement approach for spinal fusion surgery that uses virtual 3D angles between current and targeted screw trajectory, using deviation in angle to guide the surgeon (Fig. 2.11 (a) and (b)). The reported results of a lumbar spine phantom experiment indicate a promising screw insertion accuracy with the caveat that surrounding tissue was not taken into account. Other articles describing pedicle screw insertion include [122] and [99]. Percutaneous implantation of sacroiliac joint screws is presented in [89] and [90]. Catheter insertion (n=4) also has to deal with the manipulation of flexible structures and has been applied to US-guided central venous catheterisation [131], radiaton-free endovascular stenting of aortic aneurysm [132] and transcatheter procedures for structural heart disease [152]. K-wire insertion in orthopaedic procedures (n = 4) was addressed by experiments investigating fluoroscopy controlled wire insertion into femur [92], percutaneous orthopaedic surgical procedures [97] and C-arm fluoroscopy guid-

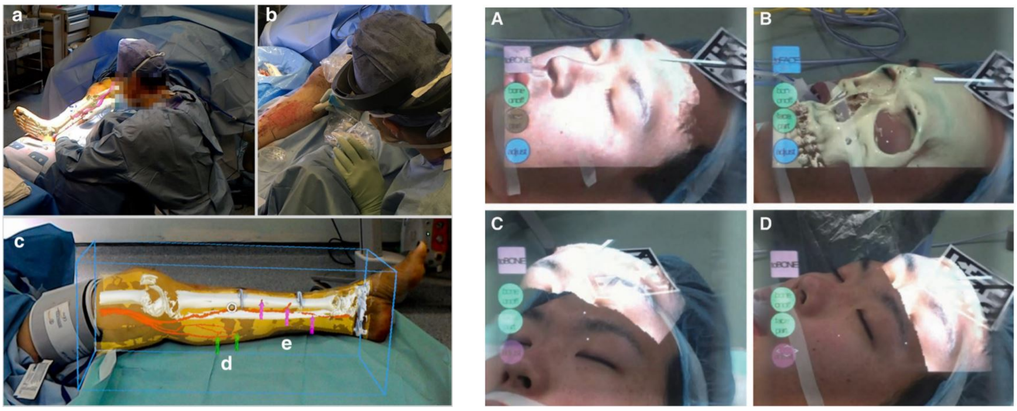


(a) (b)

Figure 2.12: (a) Robotic instrument placement and endoscopy guidance: Navigation aids for the first assistant: Real-time renderings of a robotic endoscope and robotic instruments that are superimposed on their physical counterparts. In addition, endoscopy guidance is realised via an endoscopy visualisation being registered with a viewing frustrum *(source: Fig. 4 (f) of [166])* (b) Robot placement: Reflective-AR Display aided alignment between a real robot arm and its virtual counterpart and subsequent robot placement to its intended position in preparation for robotic surgery *(source. Fig. 4 of [167])*

ance [103]. The exploration of potential benefits of 3D virtual camera views for endoscopy guidance (n = 4) has been conducted in first assistant support in robot-assisted laparoscopic surgery [166] (Fig. 2.12 (a)), percutaneous endoscopic lumbar discectomy [129] and ureteroscopy [120]. Drill trajectory guidance (n = 3) explores potential advantages of 3D virtual guidance information such as drill angle and deviation between actual and planned drill path and has been used in dental implant surgery [71] and endodontic treatments [146]. Surgical saw navigation using 3D virtual cutting guides (n = 2) was presented in mandibular resection [160] and free fibula flap harvest [105].

In addition to surgeons themselves, other clinical staff in the operating theatre can benefit from OST-HMD assitance. In minimally invasive robotic surgery it is usually the first assistant's responsibility to set up the robot arms prior to intraoperative robot control conducted by a surgeon. We identified 2 articles that present HoloLens applications aiming to support the first assistant during robot-assisted surgery: 1.) [166] robotic instrument placement in laparoscopic surgery from (Fig. 2.12 (a)) and 2.) full robot arm placement in minimally invasive gastrectomy (abdominal surgery) from [167] (Fig. 2.12 (b)). The remaining applications of surgical guidance cover topics such as stent-graft placement in endovascular aortic

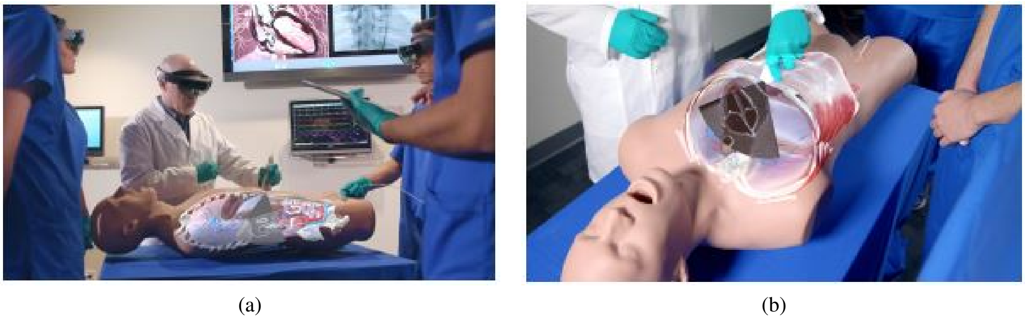


(a) (b)

Figure 2.13: (a) Dissection Guidance example in reconstructive surgery: HoloLens based identification of vascular pedunculated flaps: a Computed Tomography Angiography (CTA)-based 3D model of a female patient's leg consisting of segmented skin, bone, bone, vessels and vascular perforators lower leg is superimposed on the patient anatomy. The surgeon confirms perforator location with audible Doppler ultrasonography *(source: Fig. 3 of [158])* (b) Surgical Anatomy Assessment example in plastic surgery: AR views of the Moverio BT-200 smart glasses showing a patient with osteoma and 3D virtual facial anatomy (face surface and facial bones including the osteoma) superimposed onto a patient's face *(source: Fig. 8 of [157])*

repair [134], imaging probe navigation for tooth decay management [147], C-arm positioning guidance in percutaneous orthopaedic procedures [94], identification of spinal anatomy underneath the skin [101] and dissection guidance for vascular pedunculated flaps of the lower extremities presented by [158] (Fig. 2.13 (a)). A HoloLens based MR approach was decised in which the surgeon has to manually register a CTA-based 3D model of a patient's leg to the respective patient anatomy using HoloLens hand gesture and voice command interaction. After a surgical patient case study, surgeons confirmed that this MR solution is more reliable and less time consuming than audible Doppler US which is the conventional non-AR method.

With the main research focus being image-guidance, it is essential to consider both safety and accuracy in the design and deployment of such systems. Overreliance on the perceived precision of visual guidance, especially when not properly validated or calibrated, can introduce critical risks during surgical procedures. Additionally, excessive 3D virtual elements or poorly integrated information may con-



(a) (b)

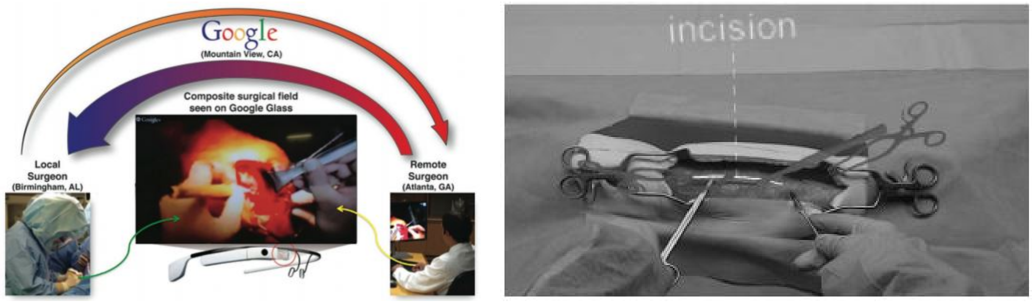
Figure 2.14: Surgical training example application: US Education. (a) Multiple users can see 3D virtual anatomical cross sections mapped on a patient simulator and the US scan plane. (b) 3D virtual subcostal four-chamber view coming out of the simulator probe. *Source: Fig. 3 and 7 of [115]*

tribute to visual clutter, potentially leading to cognitive overload for the surgeon. This increases the likelihood of errors and may also reduce trust in the system over time. Therefore, a careful balance between presenting essential information and maintaining visual clarity is required when designing such systems.

2.5.4.2 Other Surgical Application Contexts

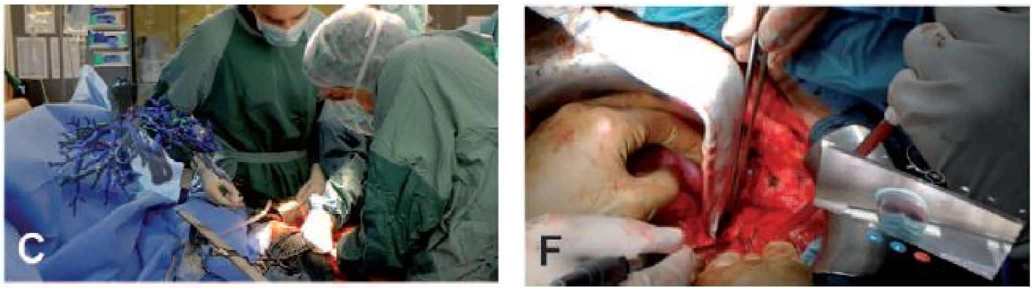
Preoperative planning applications from [150] and [149] addressed human-computer interaction issues of conventional approaches in preoperative diagnosis of coronary heart disease that lead to inaccurate diagnosis results. To overcome these limitations, the authors proposed a hand gesture-based interactive 3D virtual diagnosis system aimed at providing natural and intuitive interaction. [123] used virtual 3D vascular structures to improve the extraction and communication of complex Magnetic Resonance Imaging (MRI) image data in the context of aneurysm rupture prediction. [118] addressed planning of liver resection surgery and found that 3D virtual liver anatomy visualisations improve the user's spatial understanding. Other articles that were categorised as preoperative surgical planning investigate potential planning improvements for repair of complex congenital heart disease [151] and preoperative anatomy assessment for nephron-sparing surgery [168]. Benefits of OST-HMD AR during surgical training have been explored for preoperative diagnosis and planning of coronary heart disease [149], intraoperative surgical tool guidance during hip arthroplasty simulation [98], neurosurgical burr hole localisation [128] and transesophageal echocardiography examination from [115] shown in

Fig. 2.14.


Google
(Mountain View, CA)
Composite surgical field seen on Google Glass
Local Surgeon (Birmingham, AL)
Remote Surgeon (Atlanta, GA)
incision

(a) (b)

Figure 2.15: Telementoring applications. (a) Overview of a Google Glass systeming using a composite surgical field *Source: Fig. 3 of [88].* (b) First-person view of HoloLens-based 3D virtual instructions consisting of 3D models and 3D lines *Source: Fig. 2 of [116].*



(a) (b)

Figure 2.16: Surgical anatomy assessment and teleconsultation applications in visceral surgery: (a) Intraoperative visualisation of a preoperative model of the vascular anatomy of the cranio-ventral liver and tumor to be dissected. (b) Intraoperative tele-consulting: real-time video communication with a remote surgeon *(Source: Figure 3 (C and F) of [66])*

Telementoring also belongs to the broader scope of surgical training, but involves a surgical trainee being mentored by an expert surgeon during a surgical procedure rather than training outside the operating room. [88] used a Google Glass based mentoring system for shoulder arthroplasty (Fig. 2.15 (a)). The student surgeon and teacher surgeon can both see a composite surgical field in which hands and surgical tools of both surgeons can be seen at the same time. [116] used a HoloLens mentoring system in which an expert surgeon can place virtual 3D annotations (surgical tools and incision guidance lines) which are seen by the student surgeon in real time (Fig. 2.15 (b)). The authors reported improved information exchange between student and mentor, reduced number of focus shifts and reduced placement error. A similar mentoring is presented in [135], where trainees performed leg fasciotomies and reported an improved surgical confidence.

In contrast to telementoring, where the dialogue is continuous, teleconsultation (n = 5) focuses on a consultation based on-demand communication between colleagues. [66] explored potential benefits of using the HoloLens to establish a web-service based real-time video and audio communication with a remote colleague during visceral-surgical interventions (Fig. 2.16). In addition, the remote surgeon could mark anatomical structures within the surgical site using a tablet computer. [163] used a Google Glass for hands-free teleconsultation during different urological surgical procedures. Other examples use the Google Glass for consultation during reconstructive limb salvage [87] and orthopaedic procedures [93].

Applications where virtual 3D anatomy is displayed an intraoperative setup without trying to guide the surgical procedure are categorised as surgical anatomy assessment. This involves intraoperative assessment of preoperatively aquired patient anatomy that aids clinical decision making without trying to guide the procedure itself. [66] used a HoloLens based 3D visualisation of a liver cranio-ventral incl. tumor (Fig. 2.16 (a)) to improve a surgeon's spatial understanding of the target anatomy during dissection of the liver parenchyma in complex visceral-surgical interventions. [157] used Moverio BT-200 smart glasses and registered virtual 3D face and facial bones surfaces (Fig. 2.13 (b)) to aid clinical decision making for more objective assessment of the improvement of a patient's body surface contour in plastic surgery. A further category of display shows preoperatively acquired 2D patient imaging data and medical records in the surgeon's field of view using AR rather than a separate monitor. [163] asked surgeons to rate their perceived usefulness of displaying patients' medical records and CT scans on a Google Glass during urological surgical procedures. They found that reviewing patient images was rated less useful, whereas reviewing medical records received a high rating. [175] used a Google Glass to view and manipulate X-ray and MRI images.

2.5.4.3 AR Visualisations

Conventional computer-assisted surgery uses different types of visualisations to aid preoperative planning or intraoperative procedures and a similar range of visualisations have been adopted for AR-assisted applications (see table A4). Fig. 2.17 shows the distribution of articles by type of AR visualisation. The majority of articles use preoperative models (n = 66), usually consisting of 3D reconstructed patient anatomy generated from CT or MRI imaging content, sometimes in conjunction with preoperative planning components. [102] used 3D virtual preoperatively planned screw trajectories and drill entry points to aid pedicle screw placement in spinal fusion surgery. [158] investigated the usefulness of CT-reconstructed 3D patient leg models including bony, vascular, skin and soft tissue structures, vascular perforators and a surrounding bounding box that facilitated manual registration. Non-anatomical content, such as 3D virtual user interaction menus or graphical annotations, are also considered as a preoperative model in this review. [135], for example, used graphical annotations of incision lines and a model of surgical tools in a telementoring system. [98] implemented a virtual menu with toggle buttons for a hybrid simulator for orthopaedic open surgery training.

Applications where 3D visualisations are generated intraoperatively in order to take updated live information into account, usually for surgical guidance, we refer to as intraoperative model visualisation (n = 13). [71] used live drill trajectory guidance information such as position and depth of dental drill and injury avoidance warnings in dental implant surgery. [113] investigated utility aspects of intraoperatively generated needle visualisations such as needle position, orientation, shape and a tangential ray during needle biopsy. Live intraoperative images (n = 12) can be displayed in a surgeon's field of view using AR in order to have crucial patient data available without the need to look at a separate monitor. [96] displayed radiographic images to aid percutaneous vertebroplasty, kyphoplasty and discectomy procedures. [166] used an endoscopy visualisation in the form of a 3D plane with video streaming content that aimed to increase the first assistant's task performance in robot-assisted laparoscopic surgery. [103] explored potential benefits of 3D vir-

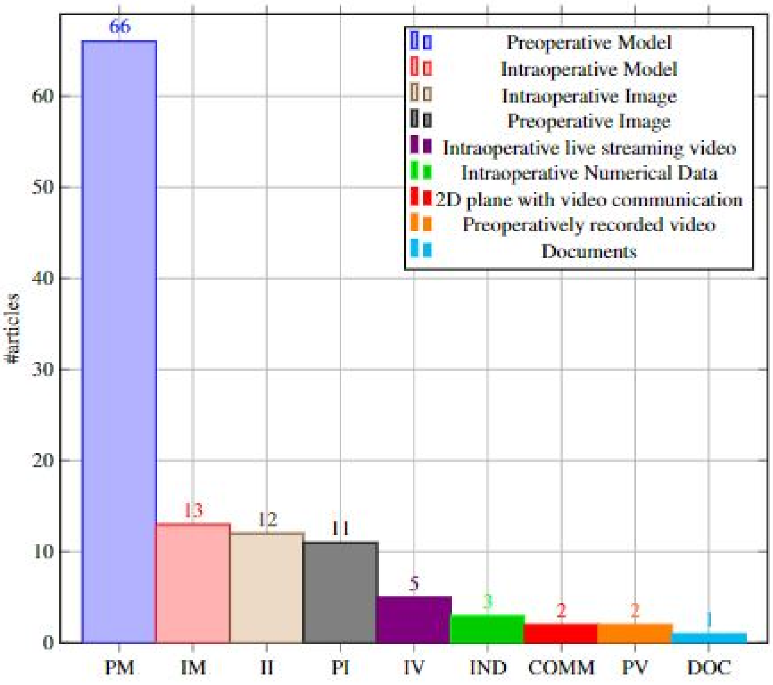


Figure 2.17: Distribution of included articles by **type of AR visualisation**

tual C-arm interventional X-ray images registered to the C-arm view frustrum for guided k-wire placement in fracture care surgery. The standard method of viewing preoperative images on a separate monitor away from the surgical site is often cited as a reason for pursuing AR guidance. 3D virtual visualisation of preoperative images (n = 5) was proposed to allow visualisation on or near the surgical site. [146] incorporated 2D radiographic images with guidance information in their HoloLens-based endodontic treatment approach. [134] used 2D images with volume rendering, arterial diameters and planning notes to support endovascular aortic repair.

The remaining categories of AR visualisations identified in this review have only been covered by a few applications. Intraoperative live video streaming (n = 5) is mostly used in telementoring applications. [88] used a hybrid image approach in which the mentee's surgical field is combined with the hands of the remote expert surgeon. [164] presented an application in which an interactive video display is visible to the mentee that shows a cursor moved by the supervising physician. Intraoperative numerical data (n = 3) is usually displayed as a 2D plane containing numerical data that aid clinical decision making or surgical guidance. [160] displayed a cutting guide deviation coordinate system supporting a surgeon during mandibular resection. [172] implemented a patient monitoring application comprising a 2D virtual screen rendered in AR that shows patient heart rate, blood pressure, blood oxygen saturation and alarm notifications. Another AR visualisation category uses a 2D plane with video communication software (n = 2) and has been applied in reconstructive limb salvage procedures [87] and orthopaedic procedures [93]. Preoperatively recorded video (n = 2) was explored by [164] as a video guide during surgical training. [87] used 3D virtual visualisation of documents (n = 1), with articles from a senior author being displayed in the surgical field of view.

2.5.5 Validation and Evaluation of AR Systems

All papers included in this review perform some kind of experiments to verify usability and the associated potential utility of their proposed OST-HMD assisted surgery solution. This section analyses the experiments conducted in each paper and categorises them as either quantitative or qualitative evaluations. An overview can be found in Appendix Chapter A, specifically in Appendix Table A4, within the *Experiments* column. The following paragraph *Evaluation methods*, briefly outlines the distinction between qualitative and quantitative evaluation approaches. The subsequent paragraph, *Experimental setting*, analyses the types of experimental environments employed accross the 91 studies included in this review.

Evaluation methods. Evaluations may consist of quantitative experiments that collect measurable data, such as registration accuracy, or qualitative experiments that gather descriptive information such as surgeons' non-measurable observations or opinions. Most of the articles in this review contain some form of quantitative experiments (n = 88), while qualitative experiments have much fewer associated articles (n = 11). Quantitative experiments include assessments of registration accuracy [89, 98, 99]), calibration accuracy [97, 103, 166]), and intraoperative guidance verification, such as tool positioning [91] or guide wire placement [102]. Experiments in which a user has to give specific survey-based feedback are also classed as quantitative. The survey is predetermined and can be evaluated numerically.

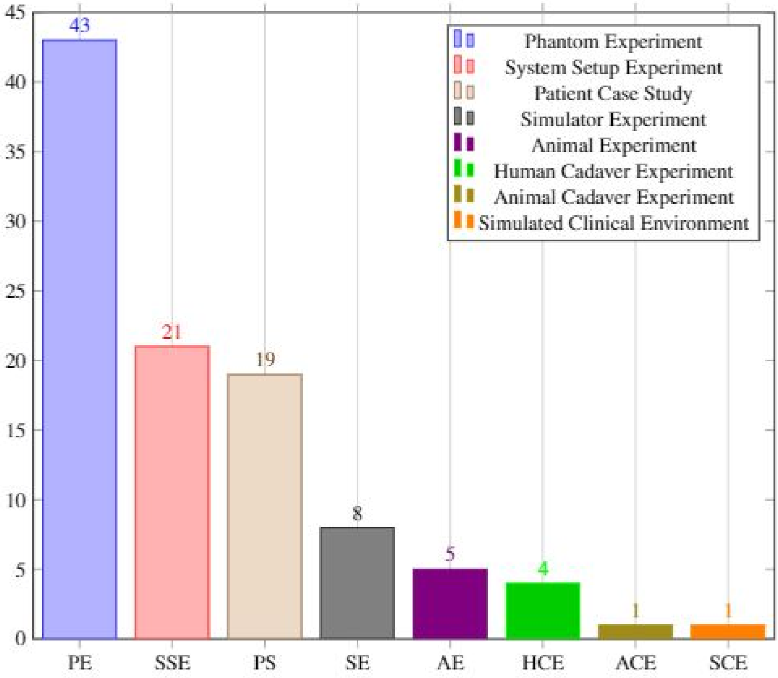


Figure 2.18: Experimental setting, from phantom to animal to clinical studies. Phantom studies dominate and though a number of clinical case studies have been reported (19), we are some way from proving clinical effectiveness of OST-HMDs at present.

Qualitative experiments are usually based on questionnaires in which participants detail specific observations that cannot be evaluated numerically. For example, [96], designed an experiment where participants completed a questionnaire following an image-guided spine surgery procedure, describing benefits, limitations and personal preferences. Both quantitative and qualitative methods are valuable and serve complementary roles in evaluating AR systems in surgical contexts.

Experimental setting. Phantom experiments dominate the list of papers (n = 43). Phantoms may be stylistic or try to mimic anatomically correct structures and are either self-made, 3D printed or acquired from specialised companies. Researchers can test their developed methods on phantoms without involving real human or animal anatomy. [89] used a 3D-printed cranio-maxillofacial model to verify the registration accuracy of their presented surgical navigation system, and a 3D pelvis model to test their navigation system. [96] incorporated a lumbar spine phantom into the validation of their presented application for image guided percutaneous spine procedures. A guidance approach for pedicle screw placement, developed by [99], was tested using a phantom consisting of L1-L3 vertebrae in opaque silicone that mimics tissue properties.

System setup experiments (n = 21) don't use realistic target anatomy structures but verify the system's intrinsic characteristics by conducting accuracy experiments in specific areas, such as system registration and calibration. [97], for example, test the calibration step of their presented OST-assited fluoroscopic x-ray guidance system that uses a multimodal fiducial. The calibration experiment consists only of a HoloLens, a C-arm and a multimodality marker. [103] conducted a similar experiment incorporating a hand-eye calibration experiment including a HoloLens, a C-arm and an optical tracker in their system that provides spatially aware surgical data visualisation. In order to verify the calibration accuracy of their proposed online calibration method for the HoloLens, [62] used a calibration box with visual markers and a computer vision-based tracking device. Patient case studies (n = 19) present surgical procedures that were tested on one or more patients. [122] validated their OST-assisted spine instrumentation approach in which neuronavigation images were streamed onto a Google Glass on 10 patients. [157] tested their intraoperative body surface improvement approach on 8 patients, each with a different diagnosis. These clinical evaluations are very useful, but further studies will be required to establish clinical effectiveness and to demonstrate improved patient outcome.

The remaining five types of experiments that have been identified in this review have a comparatively small number of associated articles. Simulator experiments (n = 8) take advantage of available simulation hardware allowing researchers or surgeons to mimic specific surgical procedures. [115] used a physical simulator model (Fig. 2.14, section 2.5.4.2) that allows users wearing a HoloLens to simulate a transesophageal echocardiography (TEE) examination. Animal experiments (n = 5) involve living animals that are anaesthetised and enable surgeons to test surgical applications under realistic conditions that consider physiological aspects such as respiratory motion. [133] and [119] tested their surgical navigation system for LDR brachytherapy on a live porcine model (Fig. 2.11 (c), section 2.5.4). [117]

performed a similar in vivo test of their respiratory liver tumor puncture navigation system that takes respiratory liver motion into account. An animal cadaver experiment (n = 1) was also performed by [71], who used a pig cadaver to test their application for intraoperative guidance in dental implant surgery. Experiments on human cadavers (n = 4) have the inherent advantage of allowing surgeons to test novel surgical procedures on real anatomic structures without posing any risk to patients. [90], for example, used six frozen cadavers with intact pelvises to investigate a novel method for insertion of percutaneous sacroiliac screws. Lastly, [93] proposed a simulated clinical environment (n = 1), aimed at testing clinical infrastructure elements and workflows rather than surgical procedures. A Google Glass based wearable personal assistant that allows surgeons to use a videconferencing application, visualise patient records and enables touchless interaction with preoperative X-ray and MRI images displayed on a separate screen without the need to use mouse or keyboard. The application was tested across various clinical setups, including a simulation doll, human actors, as well as real surgeons and nurses.

Despite the diversity of experimental settings—ranging from phantom-based setups and patient-case studies to simulated clinical environments—a clear gap remains in demonstrating consistent clinical utility, as shown by the varied experimental approaches and outcomes presented in Appendix Table A4. While many studies report promising results in terms of feasibility and usability, few provide robust clinical evidence or statistically significant improvements in surgical outcomes. In addition, the lack of standardises evaluation protocols makes it difficult to compare systems or draw generalisable conclusions. This inconsistency underscores the need for more rigoros, standardised, and clinical validation studies to support the integration of OST-HMDs into routine surgical practise.

2.5.6 Registration and Tracking in Surgical AR

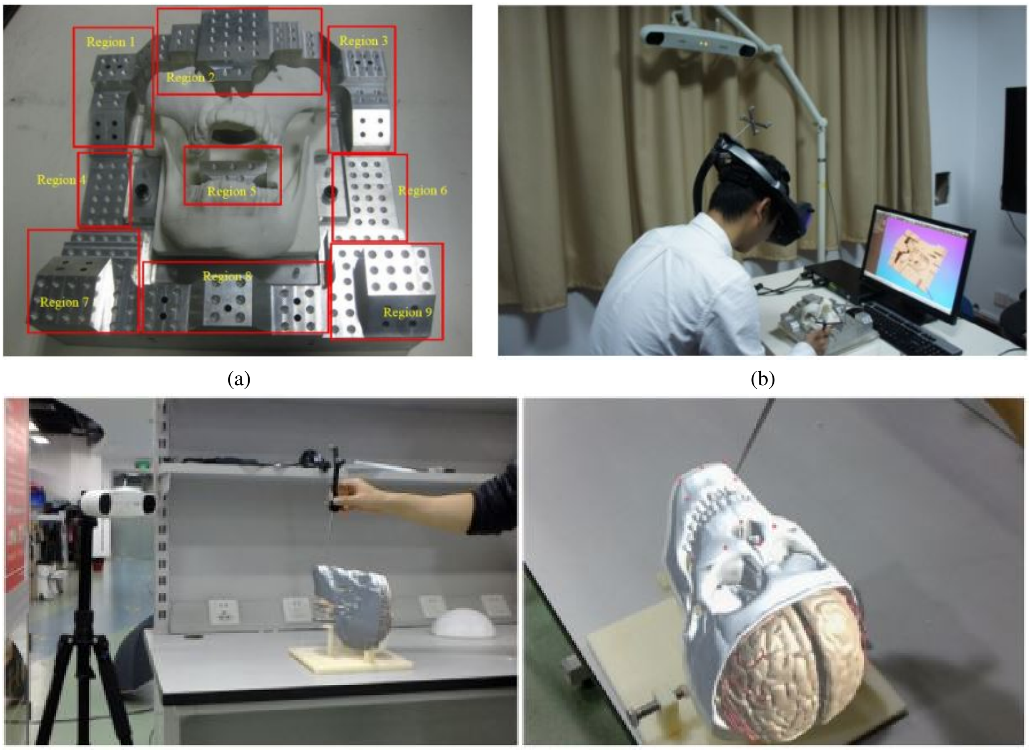
Whenever 3D virtual anatomy visualisations need to be overlaid onto a patient's corresponding anatomy, the question of registration accuracy arises. Registration refers to the establishment of a spatial alignment between the coordinate system of the patient space and the digital image space [176]. In the context of AR-guided surgery it can be defined as achieving correspondence between superimposed visualisation and patient anatomy. Devices such as the HoloLens define their own coordinate system for the room and the user's head is tracked within this space. The registration process places the preoperative model in HoloLens coordinates. If an external tracking system is used a further alignment between the devices is required. Tracking and registration each have potential errors and should be considered separately. In most cases a rigid coordinate system transformation involving translation and rotation is optimised given some corresponding features [177]. The required accuracy of the established registration depends on the application. For OST-HMD AR-guided procedures deviations between visualisation and true target anatomy may lead to surgical errors resulting from misinterpreted spatial relationships. For OST-HMD AR, overall accuracy also depends on the user's perceptual accuracy.

Appendix Table A4 in Appendix Chapter A lists the main reported accuracy results of all included articles and the associated type of conducted experiment or experiments. Because of wide variation in experimental setup and different accuracy metrics used in the literature, direct comparison of articles based on the reported accuracy is difficult. Some articles report specific registration accuracy experiments [89, 99, 117, 125, 127], while others report the accuracy of specific experimental guidance task results that result from a preceding registration [90, 91, 113, 92]. A number of papers consider manual alignment of the virtual model by the surgeon for registration. When matching corresponding features the most common methods are point-based landmark registration and surface registration [176].

2.5.6.1 Manual Alignment

[158] propose manual registration for extremity reconstruction using the HoloLens. Manual registration aligns the model directly with the HoloLens coordinate system, so no further tracking calculation is required. Since the alignment is done by the user and to their satisfaction, no correction for individual 3D perception is needed. [167] propose manual registration for virtual-to-real alignment of a robotic arm that uses two reflective AR displays (Fig. 2.12 (b)). The reflective AR displays act as

3D virtual mirrors using 3D overlays that allow the first assistant to see the virtual robot arm from multiple perspectives and therefore act as a registration aid. Experiments showed that using the reflective AR displays improved the accuracy from 30.2 ± 23.9 mm to 16.5 ± 11.0 mm. [125] compared three manual registration methods for neuronavigation using the HoloLens: tap to place, 3-point correspondence matching and keyboard control. The authors also presented a novel statistics based method allowing researchers to quantify registration accuracy for AR-assisted neuronavigation approaches. The keyboard method was found to be the most accurate (for detailed accuracy results see Appendix Table A4 in Appendix Chapter A).



(a) (b)

(c)

Figure 2.19: Accuracy verification experiment examples using optical trackers: (a) Accuracy verification block including a metal base with taper holes (for distance and angular error measuring) and 3D-printed cranio-maxillofacial model. (b) A user is conducting the accuracy verification experiment using the accuracy verification block, a tracked calibration tool and tracked OST-HMD *(source: [89]Fig. 7 and 8(c))*. Registration accuracy validation using a 3D-printed skull with 10 landmarks (red dots) and a k-wire with attached optical marker *(source: [117] Fig. 6)*

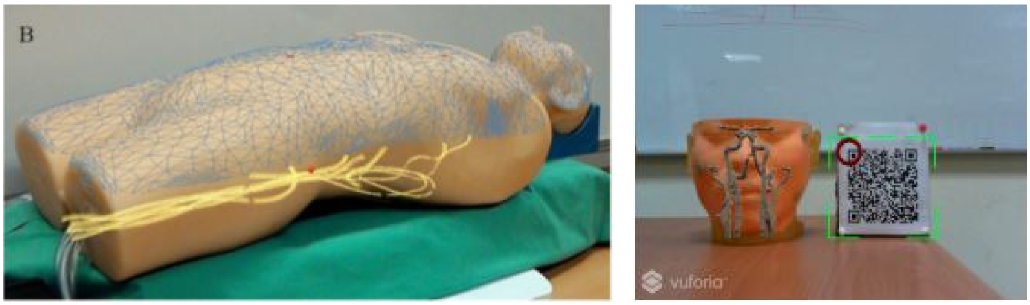
[124] presented a neuronavigation approach which is based on manual registra-

tion using fiducial markers. Users can manually register a 3D virtual visualisation of a 3D reconstructed CT scan human skull model to its physical counterpart via the help of virtual axes (Fig. 2.21 (a)). Registration accuracy was measured by both localisation accuracy (Fig. 2.21 (b)) and perceived drift of the 3D virtual overlay (Fig. 2.21 (c)). The mean perceived drift of the 3D virtual overlay during manual registration was 4.39 ± 1.29 mm. Maintaining 3D virtual object registration via continuous tracking of a marker resulted in a lower perceived 3D virtual object drift of 1.4 ± 0.67 mm.

The manual methods may not consistently achieve the level of accuracy required for high-precision surgical tasks such as dissection guidance, particularly in anatomically complex structures. In many clinical scenarios, submillimeter precision is essential to avoid critical structures, and manual alignment methods—while practical—may not achieve such required precision due to human variability and limitations in depth perception. Nonetheless, their ability to support orientation and spatial understanding remains valuable.

2.5.6.2 Point-Based Registration

Point-based registration matches corresponding pairs of fiducial points from one coordinate system to another. External fiducial markers may be attached to specific patient anatomy, such as bony structures in orthopaedic surgery or the skull in neurosurgery. Alternatively existing anatomical landmarks may be used. The same virtual fiducial points are usually marked using an external tracking device and can also be displayed on the 3D virtual anatomy model. A common accuracy measure for point-based methods is the fiducial registration error (FRE), which is the residual error of the mismatch between pairs of corresponding points after alignment. A better metric with more clinical relevance is the target registration error (TRE) at the surgical target [178]. [89] perform point-based registration as an initial alignment before surface-based refinement (section 2.5.6.3) and conducted an accuracy experiment using a verification block (Fig. 2.19 (a)) using an optical tracking system with reflective markers 2.19 (b)). The authors reported mean distance and angular errors of $0.809 \pm 0.05\ mm$ and $1.038^\circ \pm 0.05^\circ$ respectively. Addressing the problem of incorrect needle placement and associated failed tumor ablation, [117] proposed a manual registration method using a HoloLens with an optical tracker to superimpose 3D liver models on patients for liver tumor puncture navigation. Optical markers rigidly attached to the HoloLens, anatomical marks on the patient and a k-wire with attached reflective spheres serving as an optical marker are used for an initial manual registration step. The tracked k-wire is then used for automatic temporal registration during the procedure. The authors performed a registration accuracy validation experiment using a 3D-printed skull with 10 landmarks (Fig. 2.19 (c)) and reported an average target registration error of 2.24 mm. Another point-based registration approach for catheter navigation was presented by [132] and tested on a human body phantom (Fig. 2.20 (a)): A CT-reconstructed 3D body surface mesh including marching cubes segmentation of a vessel tree was registered to a body phantom using landmarks, with a reported accuracy of 4.34 ± 0.709 mm (FRE). This result highlights the difficulty of achieving accurate fiducial-based point-to-point correspondence in phantom-based studies.



(a) (b)

Figure 2.20: (a) Point-based registration: Human body surface mesh including vessel tree registered to a phantom by landmarks, where surface registration to the HoloLens surface failed *(source: [132] Fig. 1)*. (b) Surface registration result: Dummy Head with superimposed 3D CT scan reconstruction of head and intracranial vasculature. HoloLens camera detection of the QR code provides tracking *(source: [114], part of Fig. 9b)*

2.5.6.3 Surface Registration

Point-based registration is an alignment process that matches anatomical or fiducial landmarks. Surface registration offers the possibility of alignment without specific fiducial markers. Using a laser range scanner or a tracked probe, a point cloud is

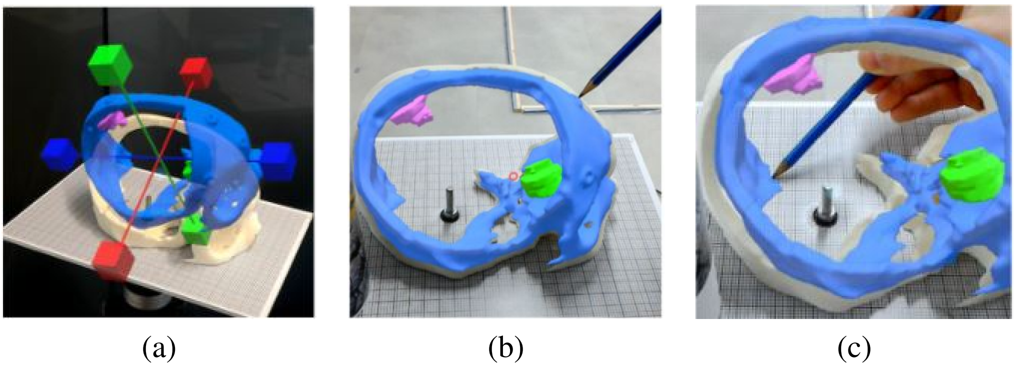

(a) (b) (c)

Figure 2.21: Registration accuracy verification using a sheet of millimeter paper: (a) Manual and point-based registration: Virtual axes allow the user to translate and rotate a human skull model in order to align it with a phantom. Fiducial markers serving as registration aids are present on both the virtual model and the phantom. (b) Localisation accuracy measurement is performed by placing the tip of a stylus into the center of a 3D virtual fiducial marker. (c) By calculating the difference in similar points the perceived 3D virtual object drift is measured. *(source: [124] Fig. 4a, 4b and 5a).*

collected from the surface of the patient's target anatomy (e.g. the head) [176]. Another surface or point cloud is derived from the image space and an algorithm is then used to match both point clouds. Most surface registration methods require a coarse manual or point-based registration step to place the image-based point cloud must be placed close to the target registration pose before the algorithm proceeds. Iterative closest point (ICP) is a popular realisation of a surface based registration and has been applied in several of our selected articles.

The HoloLens internal tracking method produces a generated surface mesh and [132] investigated whether this could be used for surface registration. A CT scan derived body surface was matched to the HoloLens surface mesh. But the HoloLens mesh resolution was found to be too coarse. In addition, [124] also reported that the HoloLens' built-in spatial mesh and simultaneous localisation and mapping (SLAM) system is unsuitable for registration and subsequent tracking due to the low vertex density and surface bias of the generated mesh and uncertainty in the SLAM realisation. [114] presented an improved version of the ICP algorithm for medical image alignment that aims to provide a global optimum via a stochastic perturbation. A dummy head alignment test revealed an average target registration error of < 3 mm. Fig. 2.20 (b) shows an example registration result.

2.5.6.4 Other Registration Methods

Other types of registration have also been explored. [152] applied a Fourier transformation based registration method in their intraoperative guidance approach for structural heart disease for transcatheter procedures. The authors used a 3D reconstructed spine image and a segmented spine from an intraoperative fluoroscopy to calculate a Fourier-based scale and rotational shift which was then used to register the fluoroscopic image to the respective 3D model of the spine. The Fourier based registration achieved an accuracy of 0.42 ± 0.02mm. A HoloLens specific markerless automatic registration method for maxillofacial surgery is presented by [161]. Their algorithm accesses the HoloLens' built-in RGB camera and extracts facial landmarks from the camera's video stream. Via known virtual-to-real world transformations of the landmarks and spatial mapping information from the HoloLens' Spatial Mapping API, the algorithm then computes the registration. The achieved average positioning error of the x, y, z axes was 3.3 ± 2.3 mm y: -4.5 ± 2.9 mm and z: -9.3 ± 6.1 mm respectively. While still at a relatively early stage, such methods suggest a shift toward more automated and context-aware registration systems that may better meet clinical accuracy requirements. However, as only two such specialised registration approaches were presented in this subsubsection, this conclusion cannot be generalised.

2.5.6.5 Tracking

Having established a registration, any subsequent motion of either the patient or the surgeon must be tracked to maintain the alignment. Tracking is a crucial component of applications such as OST-HMD-assisted surgical guidance, which require a constant alignment of 3D virtual objects with their physical counterpart. A summary of tracking methods used in the 91 included articles of this review is given in table 2.3. Consequently, the following paragraphs outline markerless tracking, marker tracking with OST-HMD device cameras, and external tracking devices.

Markerless tracking. The HoloLens inherently tracks the surgeon's head and providing the patient position is fixed within the operating room, this in itself may be a sufficient method. Eleven papers rely solely on HoloLens tracking and are

Table 2.3: Papers by tracking method

Tracking method totals		**Tracking marker totals**	
External tracker	19		
NDI Polaris	11	Reflective spheres	14
NDI EM/Aurora	4	EM	4
OptiTrack	1		
PST Base	1		
VICON	1		
Custom webcam tracker	1	Coloured catheter segments	1
Tracking with OST-HMD camera	20	**Optical Markers**	18
HoloLens	17	AprilTag	1
Other	3	Aruco	1
		ARToolkit	3
		Custom	4
		Vuforia	9
Markerless tracking	11		

associated with the manual registration process described in section 2.5.6.1. The advantage of this method is that no external measurement device is required and no markers need to be physically attached to the patient (hence the name “markerless tracking”). This can be a significant advantage in terms of sterility, convenience and operative workflow integration. However, accuracy depends on the user and may not be sufficient for some surgical tasks. [158] and [155] use manual alignment to the anatomy, while [156] register to fiducial markers for guidance of targets in the skull base.

Marker tracking with OST-HMD device cameras. OST-HMD devices such as the HoloLens incorporate cameras into their tracking process. These cameras can be used to track surface features or markers placed in the surgical field, accounting for 20 of the reviewed papers. It is common for these markers to be small planar identifiable markers modelled on QR codes. Several quite similar free libraries are available for this purpose, including Aruco, ARToolkit and AprilTags. [97] use ARToolkit markers that are also visible in X-ray to align to fluoroscopic views for orthopaedics. [102] use the stereo HoloLens camera sensors in research mode to track planar sterile markers for pedicle screw navigation. Some authors use their own custom markers, such as the cube and hexagonal markers used by [133] in their system for brachytherapy. The commercial Vuforia package can also be used to track any planar printed image and accounts for half of the marker-based tracking through the OST-HMD (9 papers). An advantage is that the position of the OST-HMD camera is relative to the surgeon, eliminating the need for an extra registration and implicitly directs the camera toward the surgical field. While this can be effective, the resolution and field of view of the cameras may not be best designed for tracking within the surgical target area.

External tracking devices. There are several commercially available devices that are able to track markers within the operating room. From table 2.3 it is clear that Northern Digital Inc. (NDI) dominate this field, with the Polaris optical tracker accounting for 11 papers and their electromagnetic tracker, Aurora, a further four papers. [117] use the Polaris for liver biopsy in the presence of breathing, whereas [132] use EM tracking for endovascular interventions. Other system are optical and account for one paper each (OptiTrack, PST Base and VICON). Except for one custom tracker based on a webcam for catheter tracking [130], all optical systems use passive reflective spherical markers. It may be invasive to attach such markers rigidly to the patient, but such methods form part of several commercial image guidance systems and this is probably the most accurate way to achieve and maintain alignment.

In summary, a range of tracking methods—both internal and external—have been applied to maintain the spatial alignment of virtual and physical objects during surgical workflows. While these technologies are critical to enabling accurate and stable AR overlays, their effectiveness in real-world surgical environments is not determined by technical precision alone. The practical success of OST-HMD-based systems also depends on how intuitively they can be used, how seamlessly they fit into existing surgical workflows, and how surgeons perceive and respond to them during procedures. These considerations belong to the category of human factors, which are essential to understanding and improving the overall usability and clinical acceptance of such systems.

2.5.7 Human Factors

OST-HMDs are wearable technological devices that enable the user to visualise and/or interact with 3D virtual objects placed within their normal view of the world.

These unfamiliar devices present a novel form of Human–Computer Interaction (HCI) and their acceptability by surgeons will depend on HCI factors. Technological aspects, such as the size of the augmented field of view or system lag during streaming of video content, can affect user acceptance. But beyond these are human factors that may vary from user to user but are crucial to the utility of a technological interaction device. They encompass perceptual, cognitive and sensory-motor aspects of human behavior that drive the design of HCI interfaces to optimise operator performance [179].

However, attempts to identify consistent generic human factors that capture basic human behavior and cognition that apply to the design of optical HCI systems has been problematic and HCI design guidelines incorporating consistent human factors have not yet been established. When addressing the negative side effects of HCI aspects only, human factors are sometimes considered as human limitations. Highlighting the aspect of human error, [180] addressed aspects of human factors and ergonomics in the operating room in general with a focus on MIS and found that most medical errors are a result of suboptimal system design causing predicable human mistakes. They also state that despite efforts made by human factors and ergonomics professionals to improve safety in the operating room for over a century, increasingly complex surgical procedures and advances in technology mean that consideration of human interaction will be required to help users cope with increasing information content. We believe that similar safety aspects of human factors also apply in OST-HMD assisted surgical applications.

2.5.7.1 Human Factors in AR

In more general non-surgical AR applications, human factors have played an important role and have been explored in the context of HCI. [181] evaluated human factors in AR in 2005 and found that apart from technological limitations, human factors are a major hurdle when it comes to translation of AR applications from laboratory prototypes into commercial products. To determine the effectiveness of AR systems requires usability verification, which led them to the following two research questions:

1. How to determine the AR user's key perceptual needs and the best methods of meeting them via an AR interface?

2. Which cognitive tasks can be solved better with AR methods than with conventional methods?

They attempt to address these two questions by conducting limited but well-designed tests aimed at providing insights into HCI design aspects that contribute to utility in perceptual and cognitive tasks. These consist of low-level perceptual tests of specific designed visualisations on the one hand and task-based tests that focus only on the well-designed part of the user interface. In [182], Livingston points out that designing cognitive tasks for usability evaluation seems to be easier than designing low-level perceptual tasks, since cognitive tasks naturally arise from the given AR application, whereas it is rather challenging to design general low-level perceptual tasks that have wider applicability. He also states that the design of a perceptual task determines how generalisable the evaluation results are beyond the specific experimental scenario and indicates that a solution may be to design general perceptual tasks that verify the usability of hardware. Finding general perceptual tasks is not always easy when hardware limitations interfere with the task design. If the effect of a hardware related feature influences a user's cognition in addition to their perception, the dependence on the perceptual task will also increase. An example for such an effect is system latency in a tracking device.

2.5.7.2 HCI Design Considerations in OST-HMD-Assisted Surgery

Although human factors in the context of HCI design considerations in OST-HMD assisted surgery are likely to be very important to the success of any system, there are a few examples of such research in the literature. In addition to the need for careful experimental design that allows a generalised result, there are also technical aspects that can be addressed to minimise unwanted human behavior when using OST-HMDs. Such technical aspects were explored by [183], who evaluated human factors in variants of the *Single point active alignment method (SPAAM)* for OST-HMD calibration that require human-computer interaction. They aimed to answer the question why calibration of OST-HMDs is challenging for users; and found that human factors have a major impact on calibration error and therefore lead to significantly different accuracy results for different users. They proposed the following guidelines for the design of OST-HMD calibration procedures:

1. Calibration should not rely on head movements only
2. The user's head should be kept stabilised by minimizing extrinsic body movements
3. Careful consideration of the data collection sequence for the left and right eye so that calibration error does not bias towards a dominant eye

[62] also described the importance of human factors in the context of OST-HMD calibration. They proposed an online calibration method for the HoloLens and concluded that the accuracy of their calibration method is difficult to measure objectively since human factors impact the overall HCI experience as well as influencing the calibration accuracy. The relationship between conscious and unconscious cognitive processes should be considered as well when considering the importance of human factors in HCI. [175] addressed the necessity to consider an egocentric interaction when designing wearable HCI systems and replaced the terms *input* and *output* with *action* and *perception*. According to the authors, an improved understanding of a human's perception, cognition and actions are necessary prerequisite when it comes to the design of a HCI system that offers better cognitive support.

2.5.7.3 Human Factors Identification

Numerous human factors were identified across the 91 included articles. Although the majority of these articles did not explicitly use the term human factors, all user-related aspects described by the authors that could impact the acceptance, utility, and performance of surgery—with or without the proposed OST-HMD solution—were considered. After grouping the user-related aspects into general terms, a total of 34 human factors were identified, as described in Table 2.4. The human factors are grouped into two categories: (i) those associated with conventional

Table 2.4: Identified Human Factors, grouped into the categories 1.) Information Perception, 2.) Cognitive Processing and 3.) Control Actions

Abbreviation	Human Factor
	Information Perception
SPATIAL_PERC	Spatial perception/awareness
INC	Inconvenience
DPPC	Missing/impaired depth perception
EYE	Individually different visual processing capabilities between dominant and non-dominant eye
COMF	Perceived comfort level when wearing OST-HMD
PER_REAL_AUG	Perception of spatial relationships between real and virtual objects
IMMR	Personal degree of perceived immersion
	Cognitive Processing
ATTN_SHIFT	Attention switch between surgical site and separate computer monitor
MM	Error-prone and cognitively demanding mental mapping of 2D image data to 3D world
SLC	Steep learning curve
EXP_OUTCOME	Influence of clinician's experience on surgical outcome
DIST	Distraction
INTPN_2D_DETAIL	Risk of incorrect interpretation of 2D image details
INTRA_OP_NAV	Impaired intraoperative navigation abilities due to absence of visual aids
COMM_3D	Personal 3D anatomical imagination capabilities affect communication between experts
CONF	Confidence
FRUS	Frustration
SUBJ_MEAS_OUTCOME	Subjective measurement of surgical outcome
EASE_HCI	Perceived degree of ease and intuitiveness of HCI
CLIN_EXP_2D	Dependence on clinical experience for interpretation of 2D image data
EMP_EST_2D	Inaccurate empirical estimation of target locations in 2D anatomy images
ANAT_PLN	Impaired anatomical understanding during preoperative planning due to 2D imaging data
CONC_LS	Loss of concentration
MIP	Limited mental information processing abilities
STRESS	Experience of stress
ENG_MOT	Engagement and motivation
PREF_HOL	Preferred degree of superimposition of 3D objects onto the surgical field (precise vs shifted superimposition)
USEF	Perceived usefulness of OST-HMD
ANX	Anxiety
	Control Actions
VIS_OPT	Selection of preferred mode of visualisation
SURG	Increased risk of surgical error
HEC	Unfamiliar/cognitively demanding hand-eye coordination
TOOL_ADJUST	Error-prone manual tool adjustment
FAT	Visual Fatigue

surgery that are addressed by OST-HMD-based AR solutions; and (ii) persistent human factors that remain unresolved or are inherent to the use of OST-HMD sytems themselves. These factors were further categorised into three phases of user interaction, as defined by the US Food and Drug Administration in the context of a medical device user interface in an operational setting [184]:

1. Information Perception (IP), where the information from the device is received by the user

2. Cognitive Processing (CP), where the information is understood and interpreted

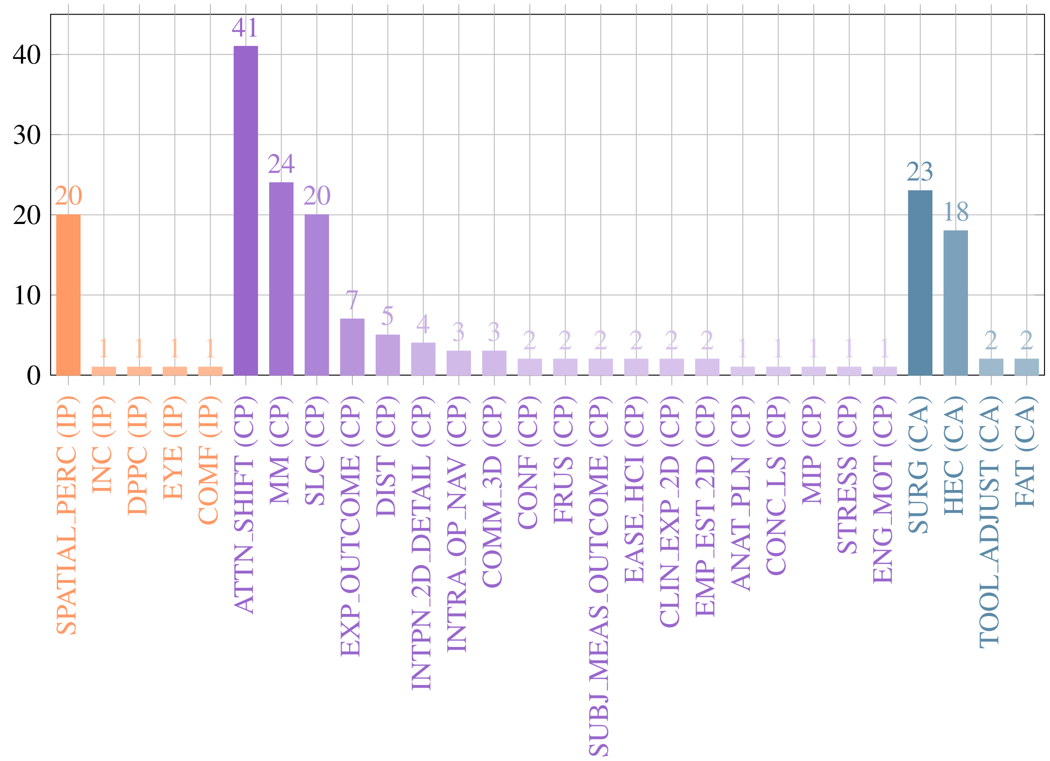


Figure 2.22: Distribution of human factors of conventional non-AR surgical approaches alleviated by the use of OST-HMD AR, grouped into the four categories 1. Information Perception (IP), 2. Cognitive Processing (CP), 3. Control Actions (CA)

3. Control Actions (CA), where this interpretation leads to actions

As a further reference, the reader is directed to Appendix Table A3 in Chapter A, which lists all addressed and persistent human factors identified in the 91 included articles.

2.5.7.4 Human Factors in Conventional Surgery Addressed by OST-HMD-Based AR

Fig. 2.22 shows the distribution of all human factors—out of the identified 34 ones—that are described as limitations of conventional, non-AR surgical methods and which the authors aimed to address with their proposed OST-HMD solution. These are grouped into the categories IP, CP and CA, according to the three phases of user interaction described in Subsubsection 2.5.7.2. The following paragraphs outline some of the most prominent human factors identified in the 91 included articles included in this review.

Information perception-related human factors of conventional surgery:

Spatial perception/awareness (SPATIAL_PERC): A fundamental limitation of conventional image guidance methods is that crucial patient anatomy can only be perceived in 2D. This restriction hinders the surgeon's ability to develop a personal sense of spatial perception and awareness. I've identified this as a key human factor in (n = 20) articles, making it the dominating human factor of conventional surgery in the IP category. Impaired spatial awareness has the unwanted side effect of an increased likelihood of surgical errors due to misinterpretation of anatomical spatial relationships. [166] aim to increase a first assistant's spatial awareness during robot-assisted laparoscopic surgery by providing a HoloLens solution. In this proposed system, a 3D virtual endoscopy visualisation is registered within the personal viewing frustrum. [103] addressed the problem of missing spatial context when looking at C-arm X-ray anatomy images on an external 2D monitor. Their proposed solution consists of a spatially aware HoloLens visualisation in which X-ray images are displayed in the correct spatial position of the patient's anatomy with a surgeon's view frustrum.

Cognitive processing-related human factors of conventional surgery:

Attention switch between surgical site and separate computer monitor (ATTN_SHIFT): The human factor with the highest number of articles (n = 41) in the CP category (and the dominating factor accross all three categories IP, CP and CA) is the *attention switch between the surgical site and a separate computer monitor*. In computer-assisted surgery a surgeon has to look away from the surgical site in order to see patient anatomy or surgical navigation information, and even switch between the surgical site and the screen multiple times during an operation. This inability to see both the surgical site and important patient anatomy or guidance information at the same time causes unwanted human behavior such as inconvenience and may also impact the continuity of the surgery [89]. Especially during image-based surgical navigation, the surgeon must constantly shift attention while manipulating surgical navigation tools, leading to unwanted side effect such as unfamiliar hand-eye coordination, distraction and loss of concentration [90].

Mental Mapping of 2D image data to the 3D World (MM): An inherent problem of conventional computer-assisted surgery is that patient imaging data and surgical navigation information is displayed in 2D. This results in the surgeon having to mentally map (or project) 2D image data onto the 3D world in order to translate the information seen on the 2D screen to the patient or surgical navigation tool. I identified *mental mapping of 2D image data to the 3D world* in (n = 24) articles. For example, [97], addressed the problem of mental mapping in the context of intraoperative guidance in percutaneous orthopaedic surgical procedures. In these procedures, the surgeon has to place tools or implants precisely under C-arm based fluoroscopic imaging. The mental projection is counterintuitive and error-prone as a result of high mental workload and mental projective simplification.

Steep learning curve (SLC): Some conventional surgical procedures, especially those related to image guidance, require surgeons to overcome a *Steep Learning Curve* (n = 20) due to the inherent complexity of the method. [113] address this problem in needle guidance procedures which require considerable learning effort. Physicians have to recover 3D information from 2D images, while the needle may cause artifacts in the images which hinder correct identification of needle tip and target. In addition, complex hand-eye coordination is required to register the 2D images seen on a separate monitor to the patient anatomy [113]. Their proposed OST-HMD AR system aims to reduce this learning curve.

Control action-related human factors of conventional surgery:

Increased risk of surgical error (SURG): Several researchers addressed the risk of surgical error (n = 23) which appears to be a common problem in some conventional image-guided procedures. [95] highlights the fact that conventional surgical navigation systems cannot observe the surgical scene and the external navigation computer monitor at the same time as being a potential problem that OST-HMD based solutions aim to solve. In another example, [146] aim to prevent or reduce errors in root canal treatments such as accidental perforation during access cavity creation.

Unfamiliar hand-eye coordination (HEC): The fact that a surgeon has to look away from the surgical site to a separate screen (see section 2.5.7.4) while simultaneously manoeuvring surgical tools in image guided navigation causes unfamiliar hand-eye coordination because the surgeon cannot see his hands while looking on the separate screen [90]. *Unfamiliar hand-eye coordination* is tackled in (n = 18) articles. [166], for example, addressed a first assistant's impaired hand-eye coordination during blind placement of robotic and hand-held instruments in conventional robot-assisted laparoscopic surgery by registering the 3D virtual endoscopy visualisation with the visualised endoscope view frustrum (see Fig. 2.12 (a) of section 2.5.4). Another example is given by [96] who mention that a fundamental problem of conventional image guided percutaneous spine lies in an indirect guidance visualisation because radiography monitors showing fluoroscopic images are not aligned with the surgical site, which in turn hinders hand-eye coordination.

2.5.7.5 Persistent Human Factors of Proposed OST-HMD Solutions

Since OST-HMDs expose the user to new and possibly unfamiliar visual perceptions, interpretations and interaction options, these devices also introduce new human factors that should be taken into account when designing effective HCI. These are referred to as persistent human factors, as they continue to pose potential challenges even when using the proposed OST-HMD-based solution. Table 2.23 shows the distribution of persistent human factors, some of which we discuss in the following sections. Analogous to section 2.5.7.4, the human factors are grouped into the categories IP, CP and CA and the most popular are detailed.

Information perception-related human factors of AR-assisted surgery:

Perceived comfort level when wearing OST-HMD (COMF): As with all HCI devices, including computers or laptops, personal comfort is one of the most important factors influencing user acceptance. Discomfort will inevitably prevent a device from becoming a routine instrument that users enjoy working with. *Perceived comfort level when wearing OST-HMD* was mentioned in (n = 15) articles, and is therefore one of the human factors that dominate the IP category. [160]

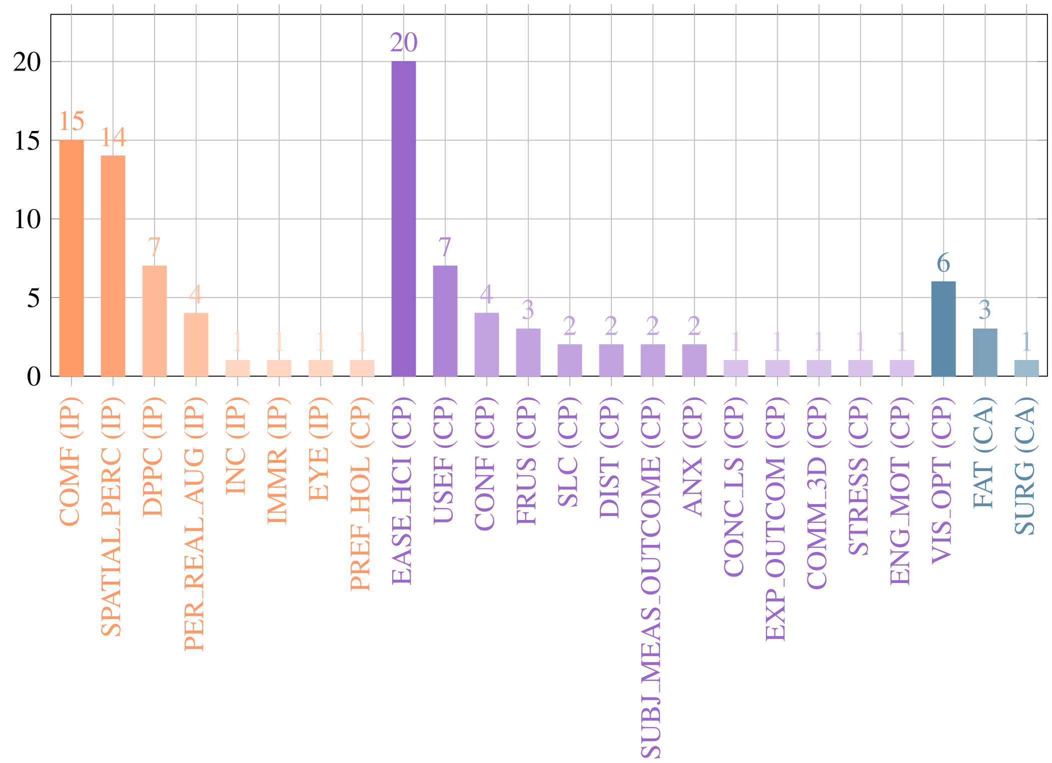


Figure 2.23: Distribution of persistent human factors of the proposed AR surgical approaches, grouped into the four categories 1. Information Perception (IP), 2. Cognitive Processing (CP), 3. Control Actions (CA)

presented a Movierio BT-200 Smart Glasses-based intraoperative navigation system that supports mandibular resection and conducted a phantom experiment in which osteotomies were performed. Surgeons reported good long-term wear work ergonomics. [135] created a HoloLens telementoring system that allows surgeons to perform mentored leg fasciotomies. Participants reported that the weight of the HoloLens has a negative impact on their posture and comfort.

Spatial perception/awareness (SPATIAL_PERC): Spatial perception and spatial awareness was already described in section 2.5.7.4 as a combined human factor of conventional non-AR methods, where 3D patient anatomy had to be inferred from 2D data. However, individual spatial processing capabilities area also factors of 3D virtual visualisations and should also be taken into account for OST-HMD solutions, as reported in (n = 14) articles. Given that AR exposes users to new perceptual stimuli that are usually not part of their normal experience, it is likely that users process this new visual information differently, which in turn impacts the quality of the HCI during OST-HMD assisted surgical procedures. [98] presented a HoloLens-based hybrid simulator for orthopaedic open surgery that allows users to visualise 3D anatomy prior to performing a virtual viewfinder-assisted surgical incision. Study participants who conducted a simulator experiment were engineers and clinicians. Results from a 5-point Likert questionnaire indicate that both user groups found it rather easy to perceive spatial relationships between real and virtual content; however, engineers tend to rate the ease of spatial relationship perception slightly higher than clinicians. Although this difference is not statistically relevant, it might be related to engineers being more familiar with 3D modelling and spatial visualisation tasks. [154] investigated in how far ear anatomy learning can be improved compared to conventional didactic lectures and computer modules. Study participants performed a spatial exploration of 3D virtual ear models displayed on a HoloLens and rated the OST-HMD higher than didactic lectures and computer modules in terms of 1.) overall learning effectiveness, 2.) the learning platform's ability to convey anatomic spatial relationships and 3.) learner engagement and motivation.

Missing/impaired depth perception (DPPC): Individual depth perception capabilities influence the ability to understand three-dimensional relationships between 3D virtual objects, as well as between real and virtual objects. This may reduce the utility of systems that require perceptual precision. Several articles indicate that missing or impaired depth perception is one of the limitations of the proposed OST-HMD approach (n = 7). [97] developed a fluoroscopic X-ray guidance system for percutaneous orthopedic surgery that is based on a co-calibration of a C arm with a HoloLens and aims to facilitate the perception of spatial relationships between patient anatomy and surgical tools. A phantom-based K-wire insertion experiment revealed that the HoloLen's build-in characteristic of rendering all 3D virtual content at a focal distance of around 2m impacts the user's depth perception and hence leads to an impaired interaction between real and virtual objects.

Cognitive processing-related human factors of AR-assisted surgery:

Perceived degree of ease and intuitiveness of HCI (EASE_HCI): A fundamental aspect that plays a pivotal role in the acceptance of a proposed OST-HMD solution is the *perceived degree of ease and intuitiveness of HCI*, which was the human factor with the most associated articles (n = 20) in the CP category. [96] presented a HoloLens based application for image guided percutaneous spine procedures that was tested in a phantom experiment in which percutaneous vertebroplasty, kyphoplasty and discectomy interventions were performed. Participants could select their preferred 3D virtual visualisation mode and questionnaire results revealed that initially the most popular mode was the option that was closest to a conventional 2D monitor and hence the most intuitive one. However, after the user became familiar with the OST-HMD environment, the preferred mode of visualisation changed to one that offers more benefits of the new mixed reality environment. [93] designed a Google Glass-based personal assistant for surgeons. Study-participants were asked to complete a post-experiment questionnaire, which revealed that some users prefer hand gesture interaction over voice interaction because voice interfered with their patient communication.

Perceived usefulness of OST-HMD (USEF): Since OST-HMDs are not yet well established in operating theaters and routine surgical procedures, clinicians can always compare OST-HMD solutions with conventional methods and thus decide for themselves whether the new AR approach is useful or not. It is therefore not surprising that *perceived usefulness* is a human factor mentioned in several articles (n = 7). [163] conducted a feasibility study of Google Glass-assisted urological procedures, in which surgeons could access 3D virtual preoperative CT scans. A patient case study with a five-point Likert scale evaluation involving 7 surgeons over 10 procedures totalling 31 procedures revealed that the system's overall usefulness was rated as very high or high by 74% of the surgeons.

Control action-related human factors of AR-assisted surgery:

Selection of preferred mode of visualisation (VIS_OPT): Sometimes users are given the possibility to optimize their HCI experience by selecting one of several visualisation modes. The *Selection of preferred mode of visualisation (VIS_OPT)* takes into account the personal optimisation of the HCI (n = 6) and is the human factor with the largest associated number of articles (n = 6) in the CA category. An example is described in [166]: a first assistant has the option between two modes of endoscopy visualisation during robotic surgery: 1.) A 3D virtual monitor capturing the endoscopy camera stream or 2.) an endoscopy visualisation that is registered with the viewing frustrum (Fig. 2.12 (a)).

2.6 Discussion

This review summarises the currently proposed applications of OST-HMDs in surgery. Orthopaedic surgery applications are the most common (30.16%) and mainly involve intraoperative guidance applications, perhaps because it involves rigid bony structures and is a field where conventional guidance systems to achieve good implant alignment have become commonplace. Image guidance, where a preoperative segmented imaging model is aligned to the operative view, is the dominating application across several surgical specialities. When providing such guidance and navigation, safety and accuracy becomes crucial. The achieved accuracy results are summarized in Section 2.5.6, highlighting considerable variation across studies. This variation is attributed to differences in experimental design and the use of diverse accuracy measures. Registration can be achieved manually or by identification of point or surface features using an external tracking device. There is no general solution to the problem of registration as yet.

The most common visualisation is of preoperative models. When these are generated from a preoperative scan this requires a segmentation process that must be incorporated into the surgical planning workflow. Medical image segmentation is a huge research area in its own right, with great progress being made. While this represents a vital component of image guidance, it falls outside the scope of this review on OST-HMD-based AR systems.

Beyond surgical guidance, this review also analyses other surgical application contexts where the accuracy of superimposed 3D virtual content in less critical, such as in preoperative planning and surgical training. Due to the variety of surgical contexts, different AR visualisations have been used, such as preoperative models, intraoperative images and intraoperative streaming of video, which all serve different purposes and are rated differently by users in terms of their usefulness. Phantom experiments dominate among the types of experiments used, underlining the fact that many such systems are some way from clinical use.

Aside from technological limitations, human factors have a major influence on the establishment of OST-HMD assisted applications in the operating room. Attention shift between the surgical site and an external computer monitor is the dominating human factor researchers aim to solve with OST-HMD solutions. These devices lead to other human factor issues, however, such as impaired hand-eye coordination and increased cognitive load that may increase rather than decrease the risk of surgical errors.

2.6.1 Classification of Human Factors

The human factors presented in this review reflect an effort to identify individual HCI-related characteristics of users in the context of OST-HMD-assisted surgery, providing an overview of perceptual and glshci-related human characteristics that may influence the utility of a proposed novel AR-assisted system. Given that OST-HMD based surgical applications have not yet replaced respective conventional state of the art methods, there is a need to raise awareness of all aspects that may influence the end user's acceptance of new technology in the operating room. Despite addressing several human factors, OST-HMD-based solutions also expose the user to new human factors that may hinder an acceptance of this novel technology in the operating room.

The dominating persistent human factor is the perceived degree of ease and intuitiveness of HCI. These new HCI possibilities may reveal individual performance differences and user preferences even more than conventional computer assisted surgical methods. Overall, it appears that the combination of OST-HMD device, surgical speciality, surgical application context, surgical procedure, proposed AR visualisation and conducted experiments triggers different individual human HCI responses that lead to variation in individual perceived utility. Some attempts have been made to provide standardised analysis in image guidance applications. [185] proposed a novel multi-indicator evaluation model for mixed reality surgical navigation systems. This model evaluates user perception in terms of safety, comfort and efficiency, combining both subjective and objective evaluation criteria. [186] identified the need for HMD-based, scientifically grounded methods that identify HCI-related interaction modalities that can be optimised to improve user performance and cognitive load. These interaction modalities comprise information presentation, user input and system feedback. Additionally, the authors suggest that an ideal HCI system should be able to adapt these interaction modalities in real-time and in response to the given task as well as environmental and user psychophysiological states.

A taxonomy for MR visualisation in image guided surgery has been proposed by [187]. The aim is to introduce a new common framework that facilitates the establishment of validation criteria and lead to more MR systems being used in daily surgical practise. The paper is well cited and the comprehensive literature review of AR in laparoscopic surgery from [79] categorises articles according to their taxonomy. But the translation into commercial applications that are used on a daily basis in operating rooms has not materialised as yet. A similar taxonomy tailored to OST-HMD AR would be desirable but is hard to achieve given the widely varying needs of the implementations presented in this review.

2.6.2 Potential Machine Learning Applications

Given the increasing trend of ML applications for medical image processing, such methods are likely to be applied to OST-HMD solutions. However, none of the selected 91 articles included in this review reported the use of ML methods in conjunction with OST-HMD systems. This highlights a potential current research gap and presents an opportunity for future investigation. OST-HMD systems inherently operate in a three-dimensional, spatially aware environment and generate a wealth of multimodal data. This includes video, hand gesture-based interaction, gaze tracking, voice commands, and real-time 3D surface meshes. These data streams offer significant potential as training input for ML algorithms, particularly related to the understanding of user behavior and respective adaptive, predictive guidance.

While none of the articles included in this literature review applied ML methods in conjunction with OST-HMD systems, related studies in surgical and non-surgical MR—outside the scope of this review—have begun to explore such approaches. [188] presented an interactive training and operation ecosystem for MR-related surgical tasks that includes data collection for potential ML algorithms. Their system records data from multiple users, such as gaze tracking to indicate which locations in 3D space a surgeon is paying attention to. ML algorithms could then use this data to identify novice surgeons and activate guidance support. Another example aiming to expand the user's hands-free interaction possibilities when wearing a HMD was proposed by [189]. Using a self-made HMD with eye-tracking cameras, the authors proposed a deep convolutional NN to classify gaze trajectories and gaze trajectory gestures. The classified gestures in turn can then trigger different HCI operations.

The HoloLens 2, equipped with eye- and hand-tracking capabilities, opens up new opportunities for such applications. However, to fully leverage these capabilities in surgical contexts, foundational work is required—particularly the development of standardised, high-quality datasets containing user interaction data. These datasets could enable the training of ML models for a wider range of purposes, including real-time registration refinement, intraoperative decision support, and personalised surgical guidance.

2.7 Summary and Conclusions

The field of OST-HMD assisted surgery has shown a significant recent upward trend in the number of publications as well as the diversity of surgical applications that could benefit from this technology. The release of the Microsoft HoloLens has notably accelerated research into MR surgical applications since 2017 (see Appendix Chapter A, table A2). However, comparatively few systems have been used clinically to date and demonstration of utility is rare. Technological limitations, such as challenges in 3D virtual object registration and perceptual accuracy, remain significant barriers to routine adoption of MR in surgery. This review also identified several human factor limitations that play a crucial role in user acceptance of such novel technology.

In particular, the way visual information is integrated into the surgical view can impact attention and decision-making. One study by [69], using a screen-based simulation system, compared direct AR with nearby, unregistered guidance. Overlaid AR was found to cause inattention blindness, where the augmented view distracts from important events in the real view. This problem arises even when registration is perfect. One option is that guidance information could be presented near to, but not overlaid directly on the surgical view. Such a side-by-side visualisation would allow correctly oriented, but not fully registered model data to be readily available without obscuring or confusing the real view.

Beyond intraoperative guidance, the potential of OST-HMD systems extends into surgical education and training. The ability to view 3D anatomy and pathology in situ may improve spatial understanding in novice surgeons and reduce the learning curve. [190] demonstrated improved learning with AR under high fidelity conditions. There is a case for similar experiments to be conducted using OST-HMD-based AR to provide evidence of proven benefit to learning with these devices. A promising research direction involves applying a human factors approach that begins by identifying specific steps or decisions within a procedure that impact patient outcome. Visualisations can then be tailored to support those tasks, which performance evaluated under controlled, lab-based conditions. This targeted strategy could both improve procedural training and clarify the optimal roles of AR in surgery. As exposure to AR technologies increase, so too may user acceptance, especially if systems are developed with both technical constraints and human factors in mind. Together, these considerations are likely to lead future research toward effective clinical implementations that realise the full potential of surgical AR.

To address the limitations identified in the reviewed literature—while continuing to explore the potential of emerging AR technologies—this thesis shifts focus toward the development of an AR-assisted medical training system. By avoiding safety-critical constraints such as technical 3D virtual object registration and perceptual precision, the work aims to investigate the educational value of MR in a controlled and lower-risk setting. The following Chapter 3 introduces a system designed to support training in obstetric sonography, highlighting the potential of MR to enhance spatial understanding and procedural learning in medical education.

Chapter 3

CAL-Tutor: Investigating the Utility of a Mixed Reality Platform for Training in Obstetric Sonography

3.1 Introduction

Addressing a relatively underexplored area in AR-assisted medical education, this chapter presents the design and evaluation of a novel MR training system developed to support clinical education in obstetric US. The CAL-Tutor system aims to enhance the spatial understanding required for interpreting 2D US images by providing interactive 3D virtual guidance in a simulated environment using a real conventional US scanner and phantom patient anatomy. The chapter outlines the educational challenges in obstetric sonography, introduces the CAL-Tutor platform, and describes the hardware and software implementation. It also reports on a preliminary user study involving novice participants, with a focus on usability, user experience, and recorded interaction data. Much of the content is based on a peer-reviewed publication by the author in the Journal of Imaging [54], and the full software implementation is openly available via GitHub[1].

Obstetric US plays a crucial role in prenatal care, enabling clinicians to assess fetal development and maternal anatomy in real time without posing harm to the

---

[1] `https://github.com/manuelbirlo/CAL-Tutor`, accessed 12 September 2024.

mother or foetus. However, learning to interpret 2D US images presents inherent challenges, particularly for novice practitioners. This section introduces the principles of conventional obstetric sonography, outlines key educational challenges, and explores the potential of MR to enhance US training. It concludes with the rationale for developing the CAL-Tutor system as a novel MR-based educational tool.

3.1.1 Importance of Obstetric Ultrasound

US is an imaging technology that plays an essential role in obstetrics, as it allows clinicians to obtain images of foetus and maternal anatomy[191], which are used to diagnose medical conditions such as fetal abnormalities and guide prenatal interventions [192]. To obtain US images, clinical practitioners must place an US transducer on the maternal anatomy. The transducer is a handheld device that emits sound waves that bounce off anatomy and generate echoes, which are received by the transducer [2]. These received echoes are then processed by computer software to generate a visual representation of the anatomy, known as a sonogram. Clinicians who perform US examinations are referred to as sonographers. Conventional transducers are connected to a larger US system equipped with a monitor and user interaction options, allowing clinicians to process and analyse the visualised anatomy. US devices that are widely used in clinical practice include the Voluson™[3] product family.

To diagnose fetal conditions, sonographers follow standardised techniques for measuring key biometric parameters, including head circumference (HC), abdominal circumference (AC), and femur length (FL) [193, 194]. Obtaining these biometric measurements requires acquiring specific US images by placing the US probe at the correct anatomical location on the mother's abdomen. These specific views are commonly referred to as US standard planes [195] or standard US planes [193]. In the remainder of this thesis, they are referred to simply as *standard planes*. An illustration of the three standard planes is depicted in Fig. 3.1.

---

[2] `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/ultrasound-transducer`, Accessed on 27 February, 2025.

[3] `https://gehealthcare-ultrasound.com/en/voluson-family/`, accessed February 27, 2025.

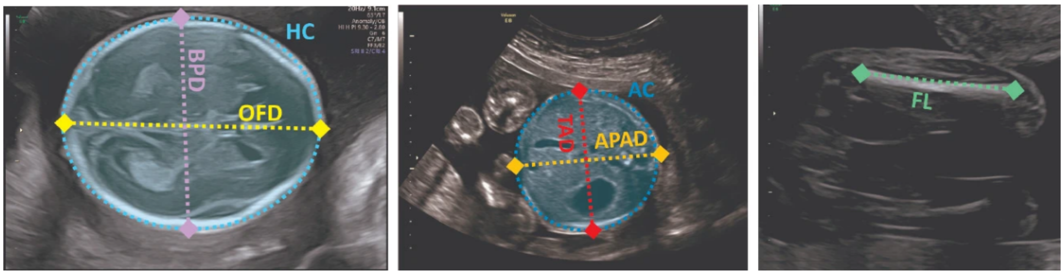


Figure 3.1: Standard ultrasound planes, used for standardisation of fetal biometry estimation. Figure reproduced from Dromey et al. [193]. From left to right: head circumference (HC), abdominal circumference (AC), and femur length (FL)

3.1.2 Educational Challenges in Obstetric Ultrasound

Obstetric US is a highly operator-dependent modality, and achieving consistent training that leads to reliable diagnostic performance remains a major challenge [196]. Sonographers must develop a strong appreciation of the US appearance of fetal anatomy in order to mentally construct a three-dimensional (3D) model of the foetus' position and structure. Since conventional US systems use a separate monitor to visualise patient anatomy, the clinician must look away from the patient while holding the transducer. Consequently, they cannot simultaneously observe the imaging results and identify the target anatomy. This is particularly problematic for novice clinicians, who must adjust their hand movements during scanning while monitoring the real-time sonogram on a separate display. This dual-tasking imposes increased cognitive load, which can ultimately reduce sonography performance [197] and pose potential risks to patient safety.

Another major difficulty is the ability to mentally map 2D US cross-sectional images onto 3D anatomical structures. While cross-sectional imaging modalities such as CT, MRI, and US provide essential diagnostic views [198], the images are inherently two-dimensional representations of intersecting planes through three-dimensional anatomy. Interpreting these requires extensive training in spatial reasoning and mental reconstruction of 3D structures [199]. In the context of fetal biometry, accurate acquisition of standard image planes—such as HC, AC, and FL—is critical for assessing fetal development [200]. Yet, consistency in these measurements is not guaranteed, especially among less experienced trainees, for whom navigating to the correct anatomical plane presents a significant challenge [201]. Despite current attempts to standardise fetal US biometry measurements through the acquisition of standard planes, persistent inter-operator variability in obtaining these measurements increases the risk of bias and errors in diagnostic data evaluation. Therefore, additional methods to improve standardisation are required [202].

3.1.3 Motivation for Using MR

The need for improved standardisation of fetal US biometry measurements motivates the exploration of MR technologies, which have already shown promise in interventional US. OST-HMD-based approaches have demonstrated transformative potential in US-guided procedures such as needle biopsies [113, 203, 204], regional anesthesia [205], and catheter placements [197], where precise spatial control is critical. The core motivation for using AR in these procedures is the ability to overlay live 3D virtual sonograms in their anatomically correct physical location–co-registered with the sound wave-emitting part of the transducer, which is part of the handheld US probe [206]. This eliminates the need to look away at an external monitor, enabling continuous visual focus on the patient. For example, a study [207] demonstrated that spatially co-locating cross-sectional images US at their correct physical position, such as on the transducer, enhances the 3D mental reconstruction of the scanned structure. This approach was found to be more effective compared to scenarios where the operator had to observe the US images on an external monitor.

More generally, OST-HMDs-assisted US aims to enhance conventional US workflows by improving visualisation, spatial orientation, procedural guidance, and training. While interventional applications remain the most prevalent use cases, their demonstrated benefits in supporting spatial reasoning suggest strong potential for use in obstetric US training. In particular, novice sonographers—-who face challenges in spatial understanding and plane acquisition-—may benefit from AR-based approaches that reduce cognitive load and provide hands-on spatial guidance [208, 209]. Therefore, the idea to leverage MR for improving the standardisation of biometry measurements in obstetric US training is a natural extension of its demonstrated benefits in other US-based contexts. This emerging evidence base motivates the development of targeted MR training systems that address the unique spatial and cognitive demands in obstetric sonography.

3.1.4 Motivation for the CAL-Tutor System

To address the aforementioned challenges in obstetric US training and to leverage the potential benefits of MR, this thesis presents a HoloLens 2-based MR application named *CAL-Tutor*. The core idea is twofold: first, to reduce a trainee's cognitive load and spatial misinterpretation by improving mental mapping from 2D US slices to 3D fetal anatomy through the use of 3D virtual visualisation; and second, to reduce variability in standard plane acquisition by providing real-time AR guidance during US probe navigation. More specifically, the CAL-Tutor system, which uses anatomically realistic 3D models of an US-compatible training phantom, assists the trainee by providing two key 3D virtual visualisations: (i) A combined view of the phantom model–based fetal anatomy – including the mother's abdomen – and the corresponding US slice, displayed in their correct physical location; (ii) Real-time MR guidance during US probe navigation toward the three standard planes — head circumference (HC), abdominal circumference (AC), and femur length (FL). In order to remain as close to conventional training in obstetric sonography, CAL-Tutor uses a conventional US scanner and commercially available phantom of mother's abdomen and foetus.

An overview of the system setup is provided in Figure 3.2. The user wears a HoloLens 2 and navigates the US probe toward a predefined 3D virtual target US plane (not visible in the figure). To track the location of the probe, a paper cube with printed ArUco markers is attached to the probe using a wooden stick. The cube is then tracked via software that uses the HoloLens' RGB camera to determine its 3D position. A detailed overview of CAL-Tutor's implementation—including the AR visualisations used to guide the user toward three distinct virtual standard US planes—is described in Section 3.3.

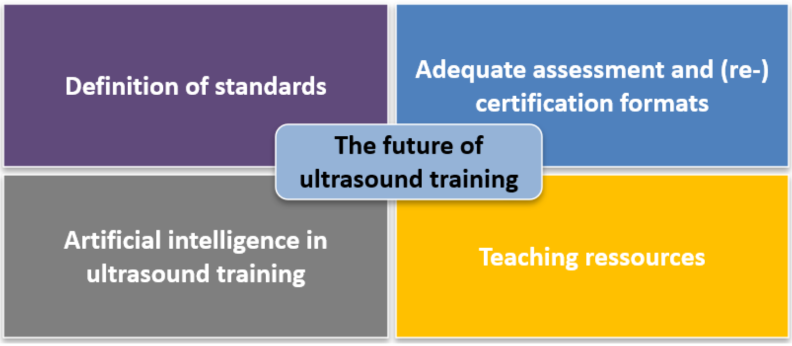
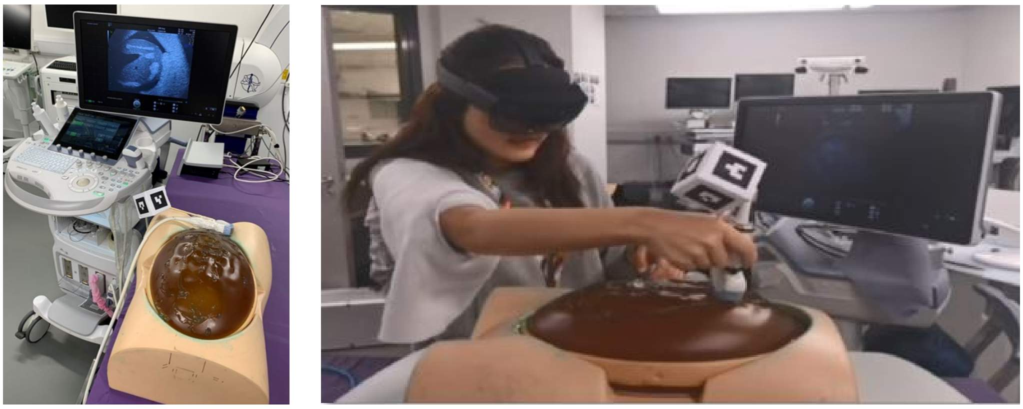

Definition of standards
Adequate assessment and (re-) certification formats
The future of ultrasound training
Artificial intelligence in ultrasound training
Teaching ressources

Figure 3.3: Potential areas to consider for innovation in future US training. Figure adapted from Wittek et al. [191].



(**a**) (**b**)

Figure 3.2: CAL-Tutor system design, showing (**a**) the setup of the US system, ArUco cube tracker and phantom; and (**b**) System in use: Navigating the tracked US probe to the 3D virtual standard plane while wearing the HoloLens 2

3.2 Background and State of the Art

To provide a broader understanding of how CAL-Tutor fits within the existing landscape of US-related medical AR, this section reviews background work and related state-of-the-art methods. At the time of the CAL-Tutor publication [54], on which this chapter is based, AR-based obstetric US training had received relatively limited attention in the literature [210]. Consequently, this section first summarises work available at the time, followed by more recent developments presented under the heading *Recent Advances in AR-Assisted US Training*.

Before delving into more technically oriented methods, it is worth noticing that researchers in obstetric and prenatal medicine also emphasise the need for innovative approaches to US training, with the ultimate goal of improving diagnostic skills and patient safety. Wittek et al. [191] identified four key areas for innovation in modern obstetric US training: (i) the definition of standards; (ii) improved assessment and certification formats to ensure consistent, high-quality training across clinicians; (iii) the integration of AI into US training to enhance quality and efficiency; and (iv) the use of technology-based training resources, including e-learning platforms, high-fidelity simulators, VR, AR, and hybrid learning platforms, to provide flexible, accessible, and cost-effective education. CAL-Tutor addresses several of the key areas for innovation in future US education identified by Wittek et al. [191]. First and foremost, it can be considered a hybrid learning platform that combines AR technology with conventional US training equipment, including a Voluson scanner and a physical phantom representing a mother's abdomen and foetus. Additionally, CAL-Tutor records user motion data, enabling the potential future development of AI-assisted feedback and guidance. Finally, the structured use of AR-based training protocols may contribute to improved standardisation in US education.

Given the importance of US probe tracking in CAL-Tutor's implementation, the next section (Section 3.2.1) reviews relevant tracking methods used in medical AR. This is followed by an overview of prior work in AR-assisted US training across obstetric and related clinical domains in Section 3.2.2. To conclude, Section 3.2.3 reviews existing DL-based standard plane navigation methods, which are more prominently represented in the research community than AR-assisted approaches.

3.2.1 Tracking Methods in Medical AR

Accurate tracking of the US probe is essential in MR applications for sonography, as it enables the correct positioning of 3D virtual sonograms at their spatially accurate physical location on the US probe. In medical AR systems, tracking can be achieved using physical markers, such as retro-reflective spheres [211], or fiducial markers like ArUco [212]. Retro-reflective spheres are tracked using infrared cameras, whereas ArUco markers are detected with conventional RGB cameras.

Electromagnetic sensors can also be used, either alone or in hybrid approaches, to address marker-related occlusion issues [213].

However, each method comes with specific limitations. Retro-reflective and ArUco markers require a direct line of sight between the marker and the tracking camera. In the context of US, the distinct shape of the probe and typical grasping positions often hinder this direct line of sight, leading to interruptions in tracking accuracy. Electromagnetic sensors, while immune to visual occlusion, are susceptible to electromagnetic interference [214, 215] and suffer from limited operational range [216], which can reduce usability. Fiducial markers like ArUco are a cost-effective solution compared to infrared or RGB-based tracking systems, but they require a fixed known location relative to the probe. This necessitates precise pre-calibration [217, 218] to ensure accurate 3D virtual overlays of sonograms relative to the wave-emitting part of the probe or to align 3D virtual probe models. Additional limitations of ArUco tracking include pose estimation noise [219], which can be caused by motion blur or non-uniform lighting [220]. Furthermore, the underlying image processing algorithms may be affected by computational load and system latency, particularly on standalone OST-HMDs such as the HoloLens 2 [47].

These challenges become even more pronounced in procedures like needle biopsies or fetal US scans, where the probe is constantly in motion. A robust tracking system must therefore account for continuous movement, while dynamically updating the 3D virtual overlay in real time [221]. However, these computational requirements place additional demands on the limited processing power of mobile AR devices and may lead to side effects such as nausea and eye strain [47]. While some of these issues have been mitigated in experimental hybrid tracking systems, their reliance on complex hardware setups and calibration limits their usability in practical training scenarios. As a result, these limitations have led to a growing interest in markerless tracking approaches. These considerations are particularly relevant to the system presented in this chapter, and their impact on design choices is discussed in Section 3.5, with further investigation of markerless tracking approaches in Chapter 4.

3.2.2 AR-Assisted US Training

The visualisation of the US plane in its physical location was one of the earliest recognised applications of AR, with the navigation of breast needle biopsy having been proposed as far back as 1996 [222]. These concepts have continued and been updated, with preliminary phantom experiments showing the potentially improved performance of biopsy needle placement [217]. Needle placement still dominates the literature in AR US guidance and training. In common with many AR systems for surgical guidance, such solutions have been proposed for some years, but remain as either lab-based experiments or small clinical studies. The lack of translation to the clinic may be due to a number of factors, including registration accuracy and as well as human factors and perceptual issues such as inattention blindness, where the augmented view obscures the visualisation of the real scene [53].

Magee et al. developed an AR-based training system for needle guidance using synthetic US imagery on a mannequin torso phantom coupled with a mock US probe [223]. A significant user study of 34 consultant interventional radiologists and 26 specialist registrars gave a favourable opinion in general, but noted that the haptic feedback was not realistic. Focusing on a low-cost US training platform, Shao et al. developed a body pose estimation-based platform for at-home skill development that only requires a printed ArUco marker attached to a simulated probe and a computer with a webcam [224]. They provided a simple system that can be conveniently used by trainees without the need for a real ultrasound machine. A user study demonstrated the utility of this concept, but revealed that the use of pre-recorded US data prevents students from learning US image optimisation. The absence of a real US probe also limits the realism of the training experience. In an attempt to offer physicians a more accessible US training solution, Costa et al. addressed the problem that conventional simulators require special hardware and developed a HoloLens application that tracks a QR code attached to a Clarius[4] wireless US probe [225]. Tracked US probe movements are then fed into simulation software that runs on an external computer, which returns an aligned US slice

---

[4] `https://clarius.com/`, accessed on 27 December 2022.

to the HoloLens at the location of the tracked QR code. A laboratory assessment of accuracy and precision showed good results, but there was no user study included.

Simulation systems offer the possibility to practice clinical skills in a controlled environment. VR training for obstetric US shows some promise and has been proposed for rehearsal before clinical training [226]. AR can also provide US simulation using video see-through devices [227]. The closest system to CAL-Tutor is the Vimedix TEE/TTE simulator (CAE Healthcare, Montreal)[5], accessed on 27 December 2022.) which offers idealised, simulated US slices visualised on 3D models from CT. The system was well received in an initial clinical evaluation [115]. While simulation has shown promise, it has not yet been adopted as a standard part of the obstetric clinical curriculum [210]. The CAL-Tutor system differs from the above training systems by using a real US scanner on a phantom coupled with AR visualisation. It is assumed that this combination of familiar, conventional hardware and novel AR technology can enhance the learning experience beyond that of idealised virtual simulation: the trainee learns the dexterity of real US, applying the right pressure while using AR to help with spatial awareness and navigation to the desired planes.

Recent advances in AR-assisted US training. In a randomised study, Farshad-Amacker et al. compared the learning outcomes of medical students when performing US-guided biopsy with and without HoloLens 2-based MR assistance [228]. Although the puncture time not differ significantly between the standard US and MR-US approach, the majority of participants found the HoloLens 2 method to be a more enjoyable learning experience. Focusing on obstetric pulsed-wave Doppler US to monitor blood flow in pregnancies, Nylund et al. proposed an MR approach to support blood flow measurement [229]. In a usability testing experiment, users with medical and IT background interacted with the proposed MR system that involves 3D-printed and visually tracked models of an abdomen and US probe. Participants found the combination of physical objects, such as 3D-printed objects, and MR visualisation beneficial for improved learning outcomes, but also noted that tracking

---

[5] `https://www.caehealthcare.com/ultrasound-simulation/vimedix/`

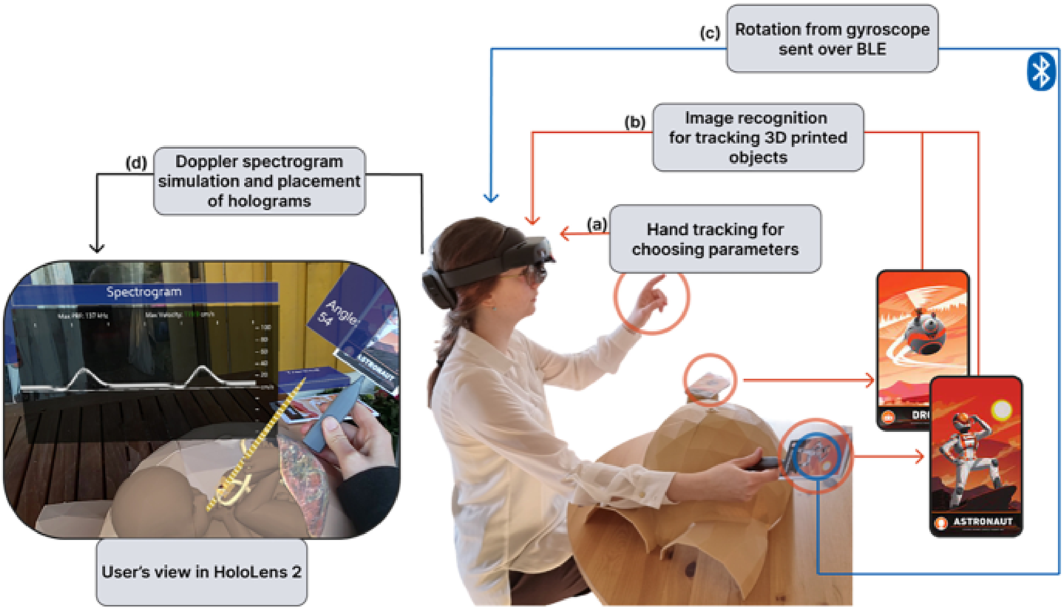

(c) Rotation from gyroscope sent over BLE
(b) Image recognition for tracking 3D printed objects
(d) Doppler spectrogram simulation and placement of holograms
(a) Hand tracking for choosing parameters
Spectrogram
Angle 54
ASTRONAUT
User's view in HoloLens 2

Figure 3.4: Overview of an MR-based obstetric pulsed-wave Doppler US training system by Nylund et al. (2024), utilising the Microsoft HoloLens 2 and 3D-printed models of an US probe and maternal abdomen. A tracked probe, along with 3D virtual visualisations of the foetus, spectrogram, and guidance information, assists trainees in simulating the measurement of fetal blood flow in the umbilical cord. Figure reused from Nylund et al. [229]

could be improved. Fig. 3.4 shows an overview of the proposed training system. Shabir et al. developed a MR tele-US system that enables an expert sonographer to guide a less experienced US operator in a remote location via a HoloLens 2 and audiovisual cues [230]. A user study confirmed that providing audiovisual cues via the proposed MR setup is sufficient for the less experienced operator to place a physical US probe at the correct target anatomy.

3.2.3 DL-Based Standard Plane Navigation Methods

Unlike CAL-Tutor, where US standard plane navigation is performed manually with MR assistance to enhance learning efficiency for novice sonographers, recent research has focused on DL-based methods for automating standard plane identification in clinical US use. These approaches aim to reduce inter-operator variability and improve workflow efficiency, particularly for less experienced users. Although such methods are not implemented in CAL-Tutor, they are relevant to this work as they offer complementary strategies that can serve as a basis for future extensions of MR-based training systems. This section briefly highlights selected examples of DL-based standard plane navigation.

Several systems integrate multimodal input data to improve standard plane guidance. Droste et al. used a conventional US system and attached an Inertial Measurement Unit (IMU) to the US probe to record freehand probe motion data. Their proposed DL network US-GuideNet consists of a Convolutional Neural Network (CNN)-Recurrent Neural Network (RNN) architecture. The CNN processes the US video signals, and the extracted image features are concatenated with the probe IMU data and fed into a RNN to predict a movement guidance signal. The network was trained with data from expert sonographers and processes both the US video signal and the simultaneously recorded motion data to predict guidance signals intended to assist the user in navigating the probe toward standard planes (Figure 3.5). Men et al. proposed a multi-modal US scanning guidance concept called Multimodal-GuideNet [231], which captures the dependency between an US video signal, the user's eye gaze and probe motion. By predicting the user's next optimal gaze position and probe movement, Multimodal-GuideNet guides users towards three biometry standard planes, the trans-ventricular plane (TVP), abdominal circumference plane (ACP) and femur standard plane (FSP). Experiments performed on obstetric scanning examinations showed that Multimodal-GuideNet outperformed single-modal deep learning models in both probe motion guidance and eye gaze movement prediction.

Other researchers have focused on learning visual features from US video streams alone. Cai et al. used a convolutional long short-term memory neural network to capture spatio-temporal visual attention information in US videos [233]. The learned visual attention maps are then used to guide standard plane detection for the HC, AC and FL standard planes. Wang et al. focused on the HC and AC standard planes only and developed a VGG16[6] [234] network-based video frame classification approach that helps operators navigate to these two standard planes

---

[6] A Convolutional Neural Network proposed by the Visual Geometry Group, Oxford University. It is now considered as outdated but is still sometimes used by researchers as a baseline model compared to newer architectures

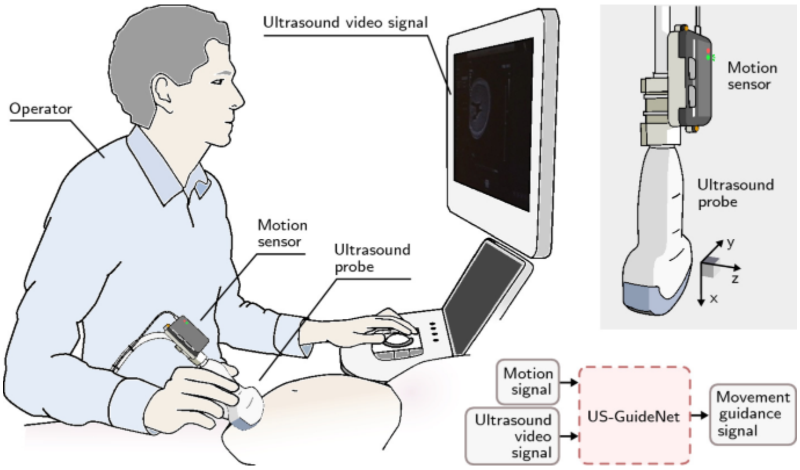


Figure 3.5: Overview of the proposed system by Droste et al. (2022) that uses a conventional US scanner and an IMU motion sensor attached to the US probe. A DL-network that processes US video signals and probe motion data captured via the IMU sensor, and predicts guidance signals for freehand probe motion. Figure reused from Droste et al. [232].

[235]. In a second step, using additional US probe motion data, they implemented an operator skill classification network. Di Vece et al., using a regression CNN, focused on the automated estimation of six-dimensional poses of US images that are in proximity of the standard transventricular (TV) plane, which is located in the fetal brain region [236]. The proposed network was trained on synthetic US images aquired from phantom-based US 3D volumens and fine-tuned on real US scans. Experimental results demonstrated good prediction performance on phantom data, but lower accuracy on real data, likely due to the domain shift between phantom and real US data. In addition, the use of a normalised brain US volume and the focus on US images around the TV standard plane only reduces the generalisability of the method.

Finally, robotic guidance systems have also been explored. Aiming to reduce the workload of sonographers and shorten US examination times, Li et al. focused on automated US scanning processes—including navigation to standard planes—performed by robotic arms using deep Reinforcement Learning (RL) [237]. A virtual RL agent, represented by a virtual US probe, operates in a 3D reconstructed US volume that depicts a virtual patient. An earlier and similar US-guided and RL-based robot navigation method, also recognised by Li et al. in [237], was presented by Hase et al. in [238]. However, this approach focused on a specific anatomical structure in the lower back (the sacrum) rather than on multiple standard planes. Experimental spine imaging tasks based on previous robot arm-based acquisitions of human patient US scan volumes show promising results, but rigid robotic arms may not lead to effective standard plane acquisitions that comply with human operator standards.

These DL approaches may provide an automated standard plane identification, which is currently performed manually in CAL-Tutor. In its current form, CAL-Tutor relies on medical experts to manually define and position virtual standard planes on the fetal anatomy. While this ensures anatomical accuracy and pedagogical control, it also limits the system's scalability and automation potential. Integrating DL-based standard plane detection methods could reduce expert dependency, enable real-time adaptive feedback during training, and support a more autonomous learning environment. As such, these methods present a valuable foundation for future enhancements of MR-based US education systems—though these extensions lie beyond the scope of this thesis.

3.3 Proposed Method

This section provides a detailed overview of the technical implementation of CAL-Tutor. It begins with the software and hardware components used in the system, followed by the design of the MR concept, the user workflow, and the data recording mechanisms. The section concludes with a description of the user study conducted to evaluate the utility of CAL-Tutor. The system was developed using a combination of hardware and software components, as summarised in Tables 3.1 and 3.2. All hardware used consisted of commercially available products: a Microsoft HoloLens 2 OST-HMD; a GE Voluson E10 US scanner, employed for conventional US training and practice; a SPACE FAN-ST phantom, a widely used modality for obstetric US training; and an NDI Aurora electromagnetic tracking system, used as a secondary modality for recording US probe motion for comparison purposes only. Similarly, the software used to implement the MR application deployed on the HoloLens 2 included only commercially available tools. Three-dimensional models of the foetus and the Voluson US probe were imported into the Unity game engine, and custom scripts written in C# were developed to enable interaction with 3D virtual objects and manage the overall application workflow.

Table 3.1: Hardware used in the implementation of the CAL-Tutor system

Hardware	**Description**
Microsoft HoloLens 2	OST-HMD (Microsoft, Redmond, Washington, USA) for spatial rendering of 3D virtual content and hand gesture-based user interaction.
GE Voluson E10	Clinical US scanner (GE Healthcare, Chicago, USA) used for standard plane training.
SPACE FAN-ST Phantom	Comercially available phantom (Kyoto Kagaku Co., Ltd, Kyoto, Japan) representing second-trimester fetal and maternal anatomy, used for training purposes.
NDI Aurora	Electromagnetic tracking system (NDI, Waterloo, Ontario, Canada) used for recording external probe motion data.

Table 3.2: Software used in the implementation of the CAL-Tutor system

Software	**Version**	**Description**
Unity Game Engine	v2020.3.14	Development platform for building the mixed reality application.
Mixed Reality Toolkit (MRTK)	v2.7.2.0	Toolkit for UI and interaction design in MR environments.
HoloLensARToolKit	v0.3	Unity-based marker tracking tool using the HoloLens 2 front-facing camera and ArUco markers [166].

The system uses two 3D virtual models derived from laser scan data: the US probe of the Voluson E10 system, and a segmented MRI model of the foetus and surrounding uterus from the SPACE FAN-ST phantom. The foetus of the SPACE FAN US phantom contains a skeletal structure, brain, four-chamber view of the heart, lungs, spleen, kidneys, aorta, UV, UA, and external genitalia. The MRI virtual model does not contain all this anatomical detail[7], but the overall anatomy of

---

[7] Since the virtual model of the phantom was reconstructed from segmented MRI data, its spatial resolution is lower than that of a model generated from a high-resolution physical laser scan.

the SPACE FAN foetus phantom is modeled in correct spatial alignment to the surface. A number of markers (Vitamin E capsules) are used for registration, which is currently manually achieved by the user. The HoloLens visualisation exhibits some instability as the user moves around, which has been noted by other authors [42]. This is probably due to the sparse scene reconstruction in the on-board SLAM algorithm of the HoloLens 2, possibly further affected by contrast-reducing lighting conditions (e.g., low ambient light and strong reflections) and by the sparsity of distinct visual features in the scene. Manual alignment by the user from their given perspective, while prone to human error, may reduce inaccuracies due to perception and head tracking. Our approach allows the trainee to train using a clinical US system, rather than a simulator device with synthetic images, as seen in many high-fidelity simulators. The standard planes (HC, AC, FL) are marked by a trainer in advance using the clinical US system. Figure 3.2a shows the system consisting of a Voluson US scanner, a SPACE FAN baby and mother's abdomen phantom, and an ArUco marker cube that is rigidly attached to the Voluson US probe by means of a wooden stick.

3.3.1 Design of the MR Concept

The central design consideration for the creation of an MR US training approach was an easy-to-use workflow that does not require advanced computer science knowledge. Therefore, all user interaction options are gathered on one 3D virtual menu that shows all available options (without hidden sub-menus) and follows the user's eye gaze but can also be pinned to remain at a fixed 3D position. However, since most of the menu buttons are for experts only, a separate toggle switch buttons disables most of these buttons so that trainees are not distracted by buttons they do not need. Figure 3.6a,b show the expert and user menu. The virtual baby model—CAL Tutor's central component, and shown in Figure 3.6c—is a 3D MRI-derived, segmented model of a foetus within the uterus, aquired from a SPACE FAN-ST training phantom. The 3D virtual model of the US probe, shown in Figure 3.6d, was reconstructed from a laser scan of a real Voluson E10 probe. A virtual cube is rigidly attached to the probe's handle at a fixed offset to indicate the position of the physical ArUco marker used for probe tracking. The virtual probe includes a dummy US plane at the transducer face, that displays a static sample US image. Live streaming of real US data is not supported in CAL-Tutor. The dummy US plane is used to instantiate copies at anatomical landmarks–defining the three standard planes for later navigation–and to aid alignment with standard planes on the baby model during navigation. Figure 3.7a illustrates the three standard US planes, placed at their respective anatomical locations of the foetus (head, abdomen, and femur) via a Unity scene. Figure 3.7b shows the basic unity components as displayed in the Unity game view. After the 3D reconstructive post-processing of MRI scans, the 3D objects of the baby and obstetric phantom as well as a laser model of the Voluson probe were imported into the Unity scene as .obj files. The objects were scaled to match the size of their physical counterparts. The US probe has an associated plane that represents the US beam produced by the probe. A 3D virtual cube with coordinate axes is rigidly attached to the probe model in a fixed distance, which is used to facilitate a user's visual confirmation of virtual to real-world alignment when the tracking of the probe is enabled. Figure 3.7b also shows the navigation components that guide the user to the target standard plane: Four pink guidance arrows originating from the corners of the US plane that is rigidly attached to the Voluson model point to the matching corners of the standard plane, and thereby serve as an additional visual guidance component that aims to facilitate the visual navigation to the standard planes. In addition to the guidance arrows, the relative distance between two US planes is displayed via position and rotation x,y,z coordinates. The elapsed time, which starts counting after the trainee presses the navigation start button on the 3D virtual menu, serves as an additional aid to make trainees aware of the time it takes them to reach the standard planes during a training session.

US probe tracking. The tracking of the Voluson US probe has been realised via a Unity asset named HoloLensARToolkit [166], which is a Universal Windows Platform (UWP) adaption of the well-known ARToolKit open source computer tracking library for AR applications. HoloLensARToolkit accesses the HoloLens's built-in webcam and tracks printed ArUco [239] markers. The *cube00-05-a4* marker, which is part of the toolkit's GitHub project, was used in this work. A printed paper version of the marker cube was attached to the US probe using a wooden stick, as shown in Figure 3.2b. A 3D virtual counterpart of the ArUco marker cube the including x, y, z, the coordinate axes and same relative dimensions is intended to help users visually confirm that the probe is being correctly tracked.

3D virtual guidance during standard plane navigation. Trainees are given several pieces of 3D virtual guidance information designed to help them navigate the US probe to one of the three standard planes, as shown in Figure 3.7b. This guidance information appears after the trainee has started the navigation phase via a 3D virtual button click and comprises the following components:

1. **Instruction card**: The card is a 2D plane with an example image of the standard plane and text explaining how to find the standard plane. The plane can be scaled and positioned anywhere in the scene via the MRTK's hand gesture-based object interaction.

2. **Guidance arrows**: Four pink arrows emanating from the edges of the US plane attached to the 3D virtual Voluson probe point to the four edges of the standard plane positioned at the respective baby location. The guide arrows are intended to enable the user to navigate to the standard planes more efficiently.

3. **Numeric offset between the source and target US plane**: The relative distance between the US plane attached to the probe and the standard plane is displayed in the upper right corner of the user's field of view via six numbers: position offset x, y, z and rotation offset x, y, z. These numbers are intended to help trainees verify that the standard plane was positioned in a precise manner.

4. **Directional indicator**: The indicator is a standard MRTK asset consisting of a chevron symbol pointing to the standard plane, helping trainees maintain a broader sense of direction when needed.

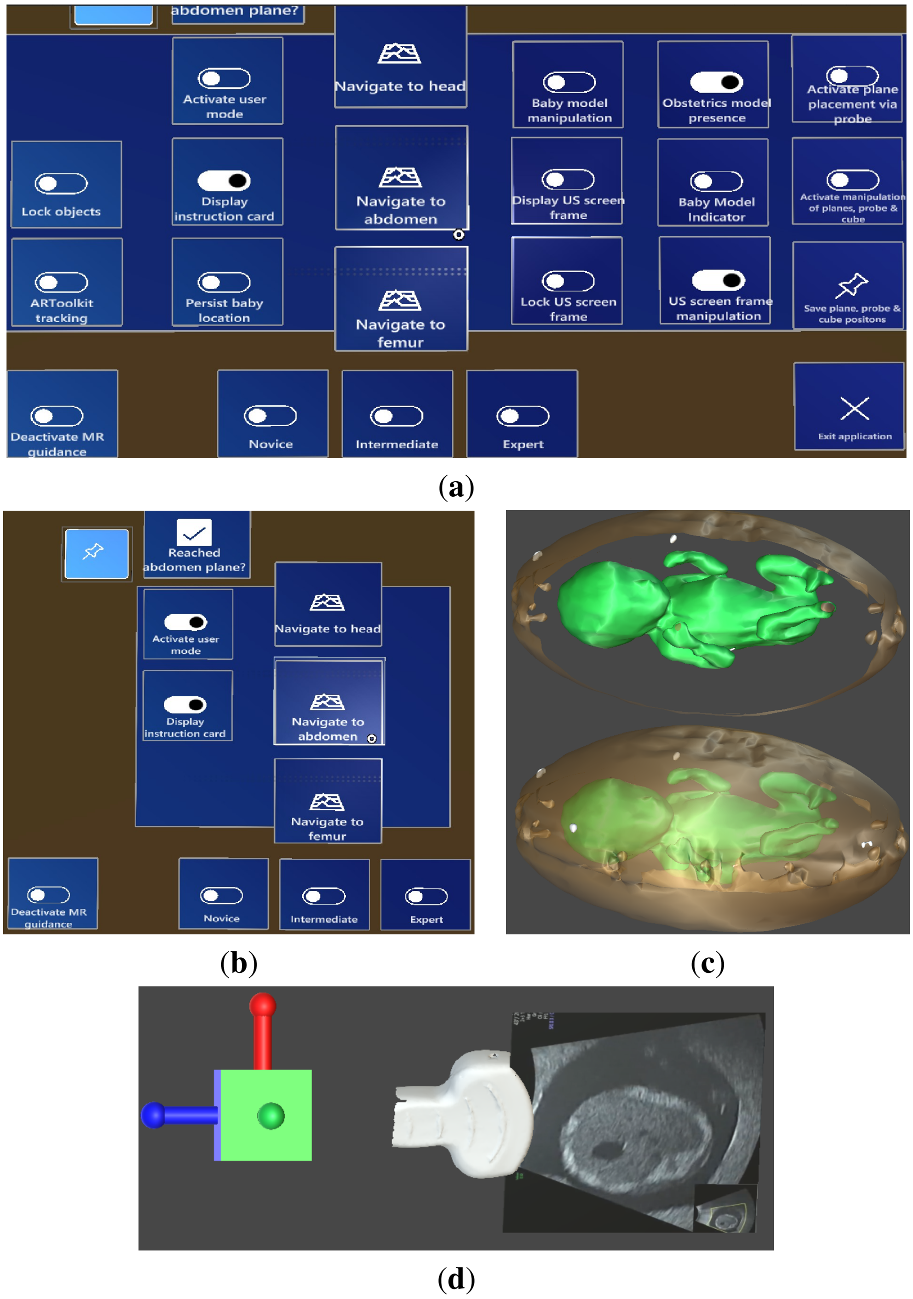
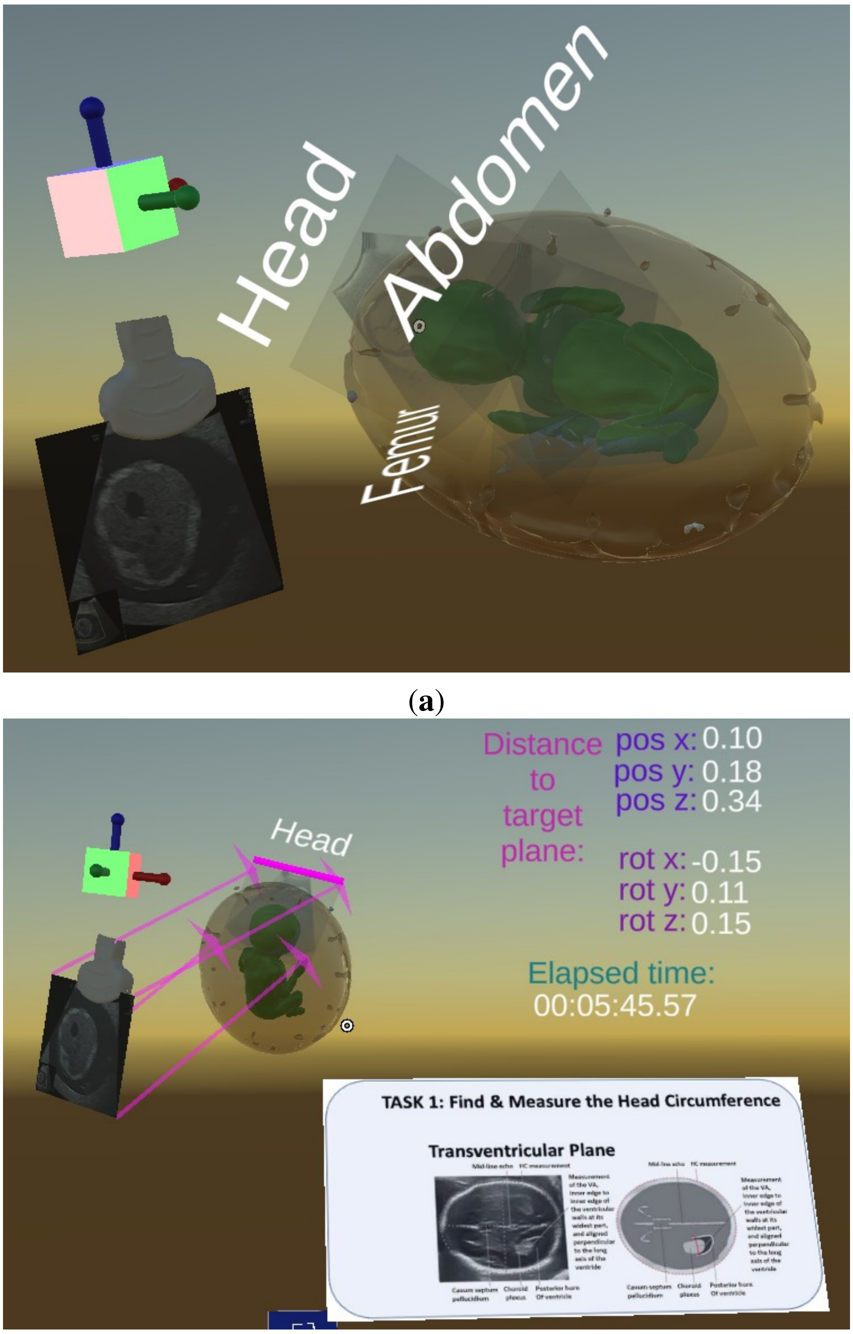


Figure 3.6: Core components of CAL-Tutor's Unity scene: (**a**) full 3D virtual menu for experts; (**b**) reduced 3D virtual menu for trainees; (**c**) Fetal model within the uterus, MRI-reconstructed from a "SPACE FAN-ST" foetus US examination phantom (top: internal view; bottom: external view through the uterine wall); (**d**) 3D model of the Voluson US probe including a dummy US plane at the transducer face and a cube indicating the position of the physical ArUco marker used for probe tracking.



(a)

Distance to target plane:
pos x: 0.10
pos y: 0.18
pos z: 0.34
rot x: -0.15
rot y: 0.11
rot z: 0.15

Elapsed time:
00:05:45.57

TASK 1: Find & Measure the Head Circumference

Transventricular Plane

(b)

Figure 3.7: Unity concepts of US probe navigation: (**a**) Standard planes (head, abdomen, femur) positioned at their anatomical locations on the baby model; (**b**) Probe navigation to the head standard plane with aids: instruction card, pink guidance arrows, numerical offset between the probe's US plane and the standard plane, and elapsed time)

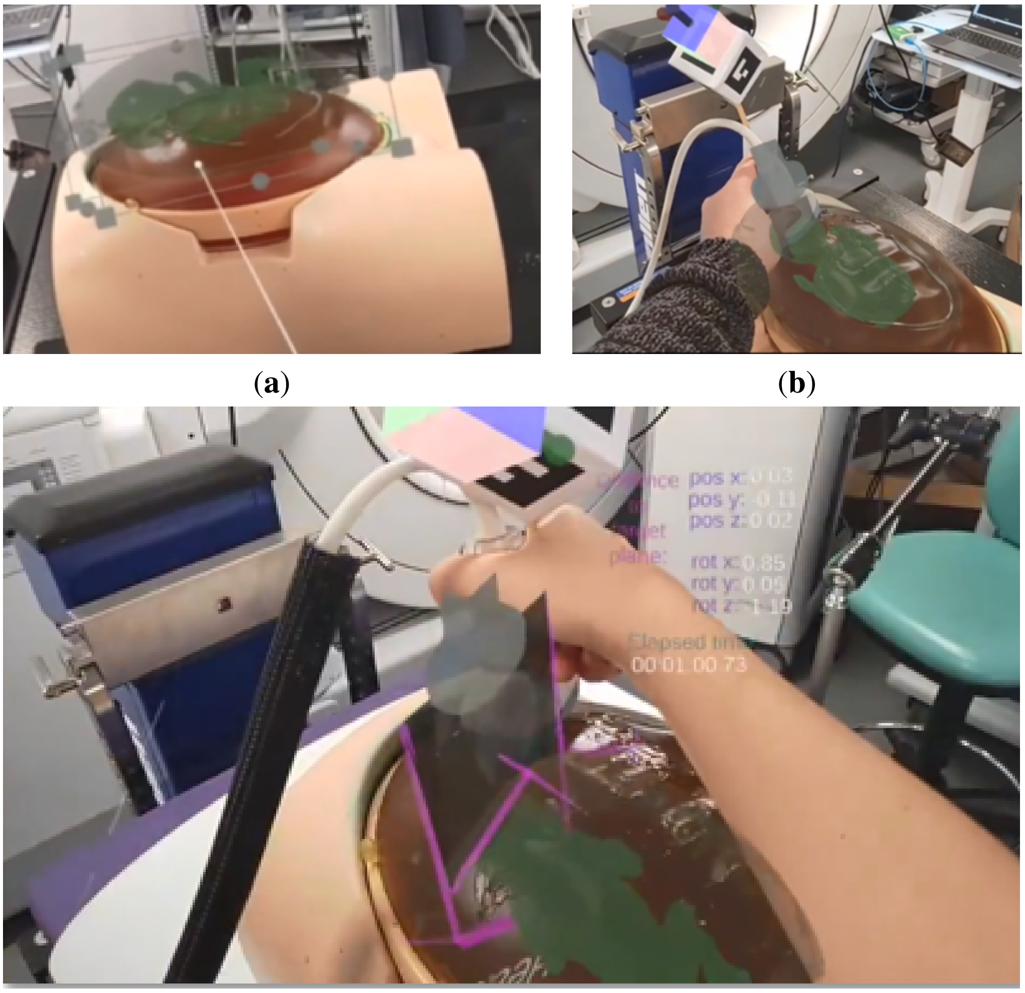
3.3.2 User Workflow

The CAL-Tutor application follows a structured, three-step workflow designed to facilitate efficient and pedagogically effective training in obsteric US. This workflow separates the tasks of expert users and novice trainees to ensure both proper application setup and educational value. Assigning initial configuration steps—such as baby model registration and standard plane placement—to experienced users helps ensure anatomical accuracy and reduces the risk of misalignment or misinterpretation. At the same time, reserving hands-on navigation tasks for trainees enables focused skill development without burdening novices with complex setup procedures. First, a medical expert aligns the virtual baby model with the physical phantom. Second, the expert defines three key standard US planes required for fetal biometry. Finally, a novice user is tasked with navigating a physical US probe to match these target planes using the guidance provided through the HoloLens 2. Each of these steps is detailed in the following subsubsections, accompanied by visual illustrations of the process. A short video demonstrating the three workflow steps of CAL-Tutor can be seen here: `https://www.youtube.com/watch?v=g0X4uLhCjoI`.

3.3.2.1 Manual Registration of the Baby Model

In the first phase of the application, a medical expert (the trainer) is expected to manually align the 3D virtual baby model to its physical counterpart via hand gesture interaction (Figure 3.8a). The 3D virtual baby model has been scaled to the actual size of the physical phantom and cannot be rescaled; only translation and rotation is allowed to manually align the model. Manual alignment of 3D virtual objects using hand gestures requires some prior familiarity of hand interaction with the HoloLens 2 and should therefore only be performed by users familiar with this device. While high registration accuracy is not expected from manual alignment of 3D virtual objects, the resulting overlay should be sufficiently accurate to provide the trainee with effective spatial orientation during the training session. As soon as the baby model was manually aligned, the experts confirmed its definitive location via a click on the 3D virtual button which freezes the model so that it cannot be moved anymore.

Optional semi-automated registration. After the publication of the CAL-Tutor system [54], a semi-automated phantom 3D virtual object registration approach



(**a**) (**b**)

(**c**)

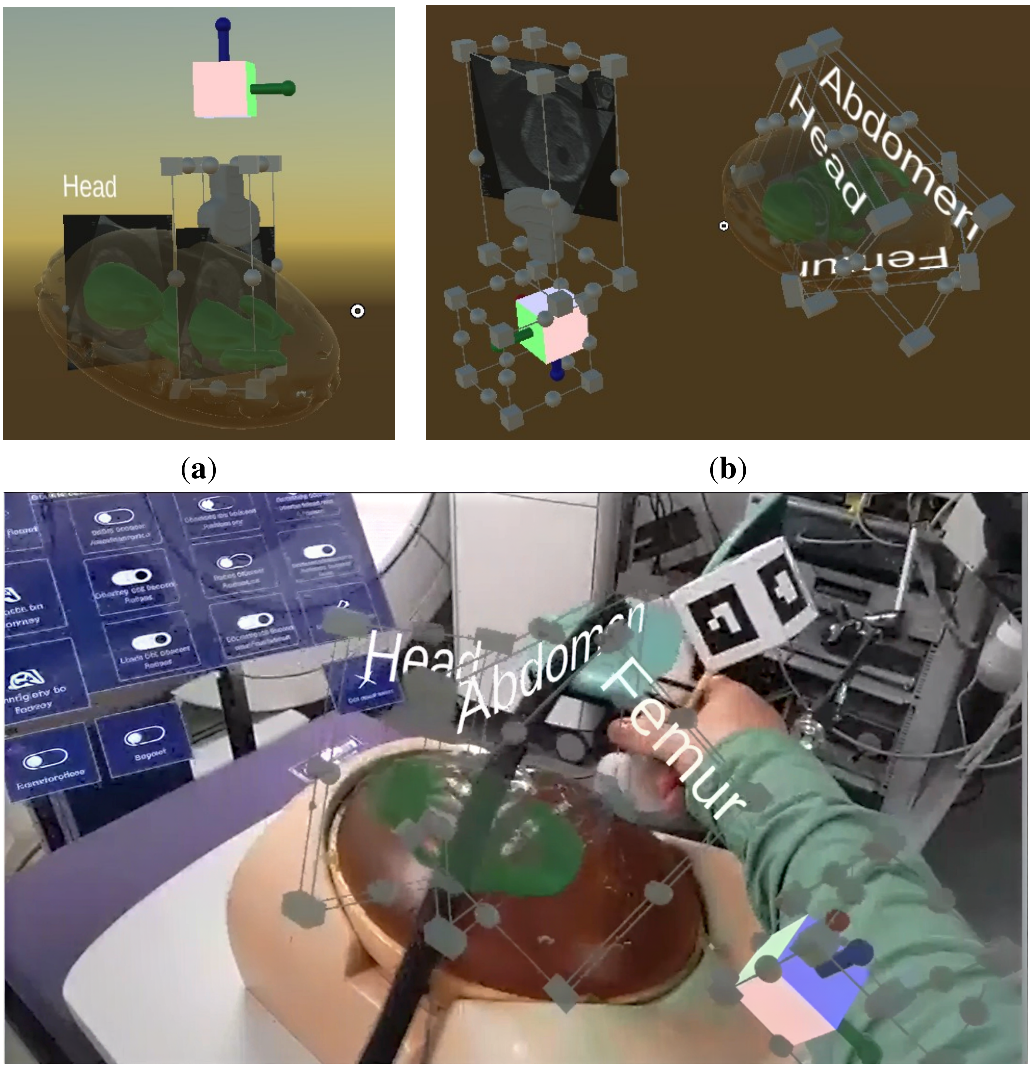
Figure 3.8: Illustration of the CAL-Tutor's user workflow phases, shown from the HoloLens 2 perspective: (**a**) The initial manual registration of the baby model (by the expert); (**b**) the manual placement of the 3D virtual standard planes at their respective baby locations (by the expert); and (**c**) trainee navigation to the standard planes.

was considered using an ArUco marker attached to the SPACE FAN phantom. The marker was tracked using the HoloLensARToolkit software. Since continuous tracking requires a permanent line of sight between the HoloLens' RGB camera and the marker, tracking is interrupted if the marker goes out of view. If the 3D virtual object has drifted in the meantime, it returns to the correct tracked position as soon as the marker is visible again. To achieve precise alignment of the 3D virtual baby model with its physical counterpart, a subsequent manual adjustment using hand gestures was still required. This ensured a consistent spatial relationship between the marker location and the 3D virtual model during usage. Although this semi-automated approach performed reasonably well, it is not part of CAL-Tutor's publicly available GitHub repository (referenced in Chapter 1, Section 1.5) and may be considered for inclusion in future iterations.

3.3.2.2 Standard Plane Definition

After the baby model was manually aligned, the expert was given two options to place the three standard planes HC, AC and FL to their respective anatomical locations of the baby model. The first option is to use the tracked ultrasound probe and place it at the respective standard plane locations with respect to the baby phantom, and place each plane individually by clicking on a 3D virtual button that creates the standard plane by taking a snapshot of the US video that is streamed onto the US plane relative to the probe. Figure 3.9a shows the concept of placing standard planes via a Unity scene: the US slice that denotes the head standard plane is placed at the respective anatomical head location and labeled accordingly, while the Voluson probe model is positioned in such a way that the next standard plane (the abdomen) can be placed. In Figure 3.8b, the manual placement of the head standard plane is shown from the HoloLens 2 perspective.

The second standard plane definition option is to use already existing standard planes positioned in their respective locations in relation to the baby model and whose x, y, z position and rotation coordinates can be loaded via a .csv file. The expert can manually position these planes via hand interaction and save the new coordinates to the .csv file. In addition, the locations of the virtual cube and US probe can be manually adjusted as well as saved in the .csv file. In Figure 3.9c, a Unity scene is shown that illustrates the concept: the three standard planes as well as the cube and probe have their MRTK-based BoundsControl and ObjectManipulator C# scripts enabled and can be manipulated. In Figure 3.9b, the manual placement of existing standard planes is shown in a laboratory setup: An expert scans the obstetrics phantom using a Volusion US scanner in order to find the exact locations of the standard planes, and then adjusts the 3D virtual standard planes accordingly via manual interaction.



(a) (b)

(c)

Figure 3.9: 3D virtual object alignment options: two options of a standard manual plane definition: (**a**) placement of new standard planes via the US probe; (**b**,**c**) adjustment of already existing standard planes whose coordinates have been loaded via a .csv file—(**b**) Unity concept and (**c**) HoloLens 2 RGB capture of the in-situ view (image upscaled for print). In this image, the background 3D virtual menu text is unreadable at this distance due to the limited resolution of the HoloLens 2 front-facing RGB camera. However, when viewed through the headset, the text is readable because 3D virtual objects are rendered directly to the displays at their native resolution and optical focus, not captured by the RGB camera.

The 3D virtual model of the US probe has an attached plane that approximates the shape and location of the US beam that is being emitted by the real probe. The virtual plane does not show the live US stream. This visualisation will not be available in clinical practice and we believe that a view of the US plane cut through the baby anatomy, where the user relates this to the ultrasound image on the scanner screen, provides effective training. After the expert reaches the location of the standard plane, the plane is pinned via a 3D virtual button click, which creates a clone of the virtual plane. In addition, the cloned US plane contains a pink bar that marks the probe-sided edge of the plane and helps trainees identify from which side they must approach the plane. A text label (‘Head’, ‘Abdomen’, ‘Femur’) helps identify the pinned plane. Once the standard planes have been placed, an unwanted shift of the virtual overlay may occur if the HoloLens performs a new spatial mapping of the scene—for example, when the device is removed and worn again. In such a case, both the baby model and the standard planes could be shifted, so that a new placement of both the baby model and the planes may be necessary. In order to facilitate a new manual alignment of the 3D virtual content, experts have the option to lock the spatial relationship between the baby model and the pinned standard planes via a 3D virtual button click. This way, only a second manual alignment of the baby model is required; the standard planes will remain at the same location relative to the baby anatomy.

3.3.2.3 Trainee Navigation to Standard Plane

In the third and last phase of the CAL-Tutor application, a trainee navigates the tracked US probe to the location of the previously pinned standard planes (Figure 3.8c). Since most of the buttons of the 3D virtual menu are only intended to be used by experts and would therefore distract trainees, a separate toggle switch button allows users to switch to an easier menu with fewer buttons (Figure 3.6a,b). Each navigation phase begins when the trainee selects the `Navigate to <target anatomy>` button—for example, `Navigate to head`. AR guidance information appears in the scene—consisting of an instruction card, guidance arrows, a numeric offset between the source and target US planes, and a directional indicator—that helps trainees find the standard planes (see Section 3.3.1). During this probe navigation phase, trainees may still look at the physical US screen of the Voluson US system in order to visually confirm that the standard plane was reached. When a trainee is confident that the standard plane has been reached, they confirm this step via a 3D virtual button click, and move on to the next standard plane.

The intention is to allow trainees to use the 3D virtual guidance information as an optional assistive tool designed to enhance their spatial understanding of fetal and maternal anatomy, while still relying on conventional US hardware for primary imaging. It is expected that some users will rely more heavily on the MR guidance, while others may continue to depend primarily on the conventional US screen. Such a difference in individual reliance on assistive MR technology may also be related to a trainee's prior experience with conventional sonography. Understanding a user's body motion during navigation of the US probe to target planes may offer valuable insights into the utility of OST-HMDs and 3D virtual guidance. Such analysis could inform future research aimed at enhancing personalised learning experiences in medical training.

3.3.3 User Data Recording

The HoloLens 2 provides a rich source of information that can be used to capture meaningful user motion data. These data can be further analysed to gain insights into user behaviour. Such motion analysis could help distinguish expert from novice interaction patterns, which may inform the development of personalised 3D virtual guidance in future MR-assisted training systems. One potential direction for future research is to leverage this motion data to train DL networks capable of predicting user behaviour and delivering adaptive MR guidance based on individual skill levels. To encourage further work in this direction and demonstrate initial feasibility, the CAL-Tutor application records user motion data during standard plane navigation sequences and stores these data in a separate `.csv` file, which can be downloaded via the HoloLens 2 device portal. The system records the US probe's position and rotation; the hand's palm and wrist position and orientation; the head's position and orientation; and eye gaze data indicating which 3D virtual object the trainee is looking at.

In addition to the HoloLens and ARToolkit data, which are accessible via the HoloLens 2 device portal, an external camera records the overall scene, and the US video feed from the Voluson US scanner is also recorded. Furthermore, an NDI Aurora electromagnetic tracker (NDI, Waterloo, Canada) was attached to the US probe to record probe motion data for comparative purposes. To enable recording of when the user is looking at the physical US screen, a 3D virtual frame can be added to the scene and manually aligned with the screen, allowing gaze data to be linked to screen attention.

Table 3.3 lists all data recorded during system use. The ARToolkit software integrated into CAL-Tutor's Unity project captures the position and orientation of the US probe. For comparison, the same data are also recorded using the NDI Aurora electromagnetic tracker. Eye gaze, hand, and head tracking data are collected through the MRTK's internal libraries. An external camera records the overall scene, capturing user interaction with both the HoloLens 2 and the Voluson scanner. In addition, the Voluson's US video feed is recorded.

Table 3.3: Data recorded by the CAL-Tutor application. The column headers denote the data source. The data listed under **NDI Aurora**, **Voluson US Scanner** and **External camera** are not available via the HoloLens device portal and need to be captured separately.

ARToolkit	**HoloLens 2**	**Voluson US Scanner**
ProbePosition$_{\text{x, y, z}}$	EyeGaze Target Position$_{\text{x, y, z}}$	US video
ProbeRotation$_{\text{x, y, z}}$	EyeGaze Target Name	
	HandPalmPosition$_{\text{x, y, z}}$	**External camera**
NDI Aurora	HandWristPosition$_{\text{x, y, z}}$	External camera
ProbePosition$_{\text{x, y, z}}$	HeadPosition$_{\text{x, y, z}}$	video of the
ProbeRotation$_{\text{x, y, z}}$	HeadRotation$_{\text{x, y, z}}$	overall scene

3.3.4 User Study

In order to investigate the potential benefits of MR guidance during US probe navigation, a small questionnaire-based user study[8] with six engineering students was conducted to evaluate the users' personal impression of the CAL-Tutor system's usability. None of the study participants had prior clinical education in obstetric

---

[8] The study was conducted in accordance with the Declaration of Helsinki and was approved by the UCLIC Research Ethics Committee at UCL (Approval ID: UCLIC/1819/006/BlandfordProgrammeEthics).

sonography. While prior experience with the HoloLens 2 was not systematically assessed, informal observation suggested that most participants were unfamiliar with the device. Consequently, residual differences in device familiarity may have influenced workload and user experience ratings. Future studies should record prior Hololens 2 exposure.

Two questionnaires were completed after the navigation tasks: (i) a NASA Task Load Index (TLX)-based workload assessment, using five seven-point scales with 21 graduations (from very low to very high); and (ii) a product assessment (user experience), using twenty six seven-point scales evaluating various product characteristics. In addition to the two questionnaires, the participants left personal qualitative notes regarding their experience with the CAL-Tutor system.

Using a commercially available Voluson US scanner and a SPACE FAN trainer phantom, the CAL-Tutor application guided users to the three standard planes: HC, AC and FL. Before the study began, a researcher manually aligned the 3D virtual baby model with the phantom and positioned the three standard planes at their respective locations relative to the model. Each participant received instructions on how to use the CAL-Tutor application and how to use the 3D virtual menu to complete the standard plane navigation sequences when wearing the HoloLens 2. After the introduction and 3D virtual model setup, participants were asked to put on the HoloLens device and perform three separate US navigation tasks—targeting the HC, AC and FL standard planes—in that specific order. This order reflects a pragmatic clinical workflow for second-trimester biometry, grouping of US planes with similar probe orientation: HC and AC are acquired in axial planes, wheres FL requires aligning the femur in a longitudinal plane. Because task order was fixed rather than counterbalanced, carry-over effects (e.g., learning or fatigue) cannot be excluded. Future studies should counterbalance task order. Each participant completed the task twice: once under the baseline condition without MR guidance (Condition 1), and once with MR guidance (Condition 2), as described in Table 3.4. The order of conditions was randomised. Nevertheless, as with the fixed sequence of standard-plane acquisition, residual learning or fatigue effects between conditions cannot be excluded. Future studies should explicitly counterbalance condition order

Table 3.4: The two experimental conditions of the user study.

Experimental Condition
Condition 1 (Baseline): Probe navigation without MR assistance The participant has to wear the HoloLens 2 device during standard plane navigation since user data will be recorded. Although the user wears the HoloLens 2, no 3D virtual information is being displayed.
Condition 2 (MR guidance): Probe navigation with MR assistance The user is asked to perform the standard plane navigation with MR guidance which includes the instruction card, the guidance arrows, directional indicator, elapsed time and numerical offset between the probe's US plane and the target standard plane, as described in Section 3.3.1.

3.4 Results

This section presents the results obtained from the evaluation of the CAL-Tutor system, as described in Section 3.3.4. The results are divided into two categories: subjective assessments (workload and product evaluation), and a preliminary objective analysis based on recorded user motion data. The subjective assessments evaluate how study participants perceived the system's usability and workload, including both task-related effort and general user experience. These are presented in Subsections 3.4.1 and 3.4.2. Additionally, to demonstrate the potential of the recorded user motion data for objective performance evaluation, an exploratory analysis of a subset of this data is detailed in Subsection 3.4.3.

Although the sample size was small and the effects were not statistically significant, the following results provide preliminary insights into user interaction and the potential role of MR guidance in standard US plane navigation training. They also aim to provide initial insights into the feasibility and effectiveness of the CAL-Tutor system in supporting standard US plane acquisition during fetal US training. Although these results are not conclusive due to the small sample size, they offer valuable guidance for future system iterations and experimental designs.

Table 3.5: NASA-TLX questionnaire results for the six study participants (engineering students) in two conditions: With MR guidance (A) and without MR guidance (B).

User Number	Condition	Mental	Physical	Temporal	Performance	Effort	Frustration	Mean
1	A	40	30	50	50	80	20	45
1	B	70	60	80	50	95	80	73
2	A	40	65	5	5	10	10	23
2	B	75	90	40	25	60	30	53
3	A	45	35	25	75	45	30	43
3	B	65	25	50	55	60	50	51
4	A	25	25	30	35	20	20	26
4	B	60	35	50	65	65	45	53
5	A	75	40	35	60	75	30	53
5	B	65	40	25	30	80	15	43
6	A	95	40	55	40	70	75	63
6	B	35	25	30	25	20	40	29

Workload	Value	Workload Component	Mean (with MR)	Mean (without MR)
Low	0–9	Mental	53	62
Medium	10–29	Physical	39	46
Somewhat high	30–49	Temporal	33	46
High	50–79	Performance	44	42
Very high	80–100	Effort	50	63
		Frustration	31	43

3.4.1 Workload Assessment

The result of the NASA-TLX workload assessment is shown in Table 3.5, which presents the individual results for both study conditions A (with MR guidance) and B (without MR guidance), as well as the mean workload values for each workload component. Even though the results are similar for Conditions A and B, the mean values are slightly better for Condition A, which indicates that MR guidance has a generally positive impact on novice users in terms of finding the three standard planes without prior obstetric US experience. Figure 3.10 shows the workload distribution generated from the data depicted in Table 3.5 which visually confirms the slightly better results for condition A.

In terms of individual workload components, notably lower values were observed in all components under Condition A, expect for *Performance*. This component is interpreted in reverse (i.e., lower scores indicate better perceived task accomplishment) and showed a slightly higher mean in Condition A than in Condition B (44 vs. 42). The slightly lower perceived performance in Condition A may be attributed to participant's unfamiliarity with the MR setup. Nonetheless, the consistently lower scores in the remaining components under Condition A suggest

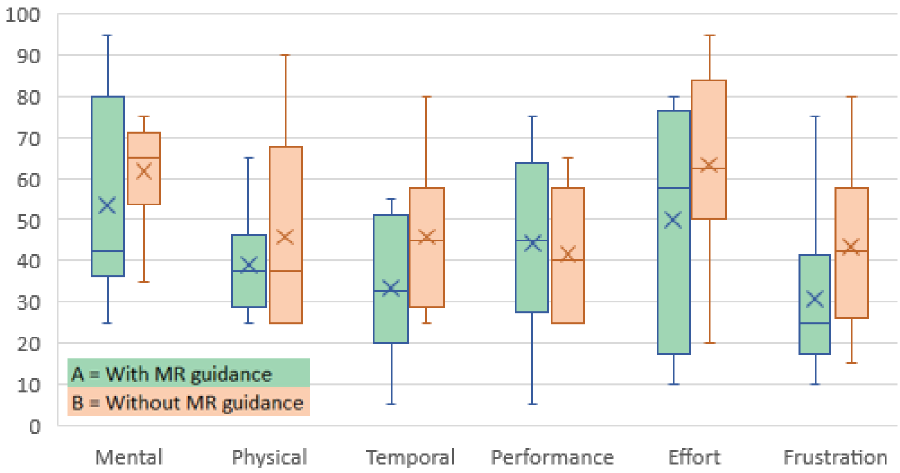


Figure 3.10: NASA-TLX workload assessment result, represented as a box and whisker chart, grouped into the two experimental conditions A = with MR guidance and B = without MR guidance.

that MR guidance has the potential to improve perceived workload in this training scenario.

3.4.2 Product Assessment

In addition to the workload assessment presented in Section 3.4.1, a product assessment was conducted to evaluate participant's subjective experience with the CAL-Tutor system. Each participant rated six product-related characterstics—*Attractiveness*, *Perspicuity*, *Efficiency*, *Dependability*, *Stimulation*, and *Novelty*—using a 7-point Likert scale, ranging from -3 (completely disagree) to $+3$ (completely agree). For each product scale, the average score per participant was calculated by averaging their responses across all items associated with that scale. The resulting individual means were then aggregated to calculate the group mean score, standard deviation (STD), and 95% confidence intervals separately for Conditions A (with MR guidance) and B (without MR guidance).

The 95% confidence interval (CI) was calculated for each scale using the formula:

$$\text{CI} = \bar{x} \pm t \cdot \frac{s}{\sqrt{n}},$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the number of participants (here, $n = 6$), and t is the critical value from the t-distribution (for

Table 3.6: Individual product assessment result of all six study participants in two conditions: with MR guidance (A) and Without MR guidance (B). CI = Confidence Interval.

Scale	**Condition**	**Mean**	**STD**	**N**	**95% CI ±**	**95% CI (Lower)**	**95% CI (Upper)**
Attractiveness	A	2.00	0.85	6	0.68	1.32	2.68
	B	0.47	0.68	6	0.54	−0.07	1.02
Perspicuity	A	1.79	0.86	6	0.69	1.11	2.48
	B	−0.42	1.37	6	1.09	−1.51	0.68
Efficiency	A	1.88	0.59	6	0.47	1.41	2.34
	B	0.29	0.95	6	0.76	−0.47	1.06
Dependability	A	1.71	0.87	6	0.70	1.01	2.41
	B	0.50	1.00	6	0.80	−0.30	1.30
Stimulation	A	2.13	0.68	6	0.55	1.58	2.67
	B	0.96	0.89	6	0.71	0.25	1.67
Novelty	A	2.21	0.25	6	0.20	2.01	2.41
	B	−0.13	2.08	6	1.66	−1.79	1.54

degrees of freedom $df = n - 1 = 5$, critical t-value $t_{0.975} \approx 2.571$). The last three columns in Table 3.6 represent: (i) **95% CI ±** – the margin of error from the mean (i.e., $t \cdot \frac{s}{\sqrt{n}}$), (ii) **95% CI (Lower)** – the lower bound of the confidence interval ($\bar{x} - \text{CI}$); and **95% CI (Upper)** – the upper bound of the confidence interval ($\bar{x} + \text{CI}$).

As shown in Table 3.6 and the box-and-whisker plot in Figure 3.11, mean scores were higher with MR guidance (Condition A) on four of the six product scales. For “Dependability” and “Stimulation”, the 95% confidence intervals overlapped, so no clear preference can be concluded. Given the small sample size ($n = 6$), no statistical differences can be inferred. Nevertheless, the pattern of higher means suggests that participants tended to evaluate CAL-Tutor more positively when MR guidance was provided.

3.4.3 HoloLens 2 User Motion Data

Even though an in-depth evaluation of the HoloLens 2 user motion data was beyond the scope of this study, Figure 3.12 presents an example of how such data could be used to further analyse user behaviour. Among all the types of user motion data recorded by the HoloLens 2, eye gaze was selected to demonstrate the potential value of such evaluations. Since clinicians must divide their visual attention between the US screen and the patient’s (mother and fetal) anatomy, analysing visual attention is particularly relevant for assessing user behavior in this context.

The visual attention profiles of the study participants during the three MR

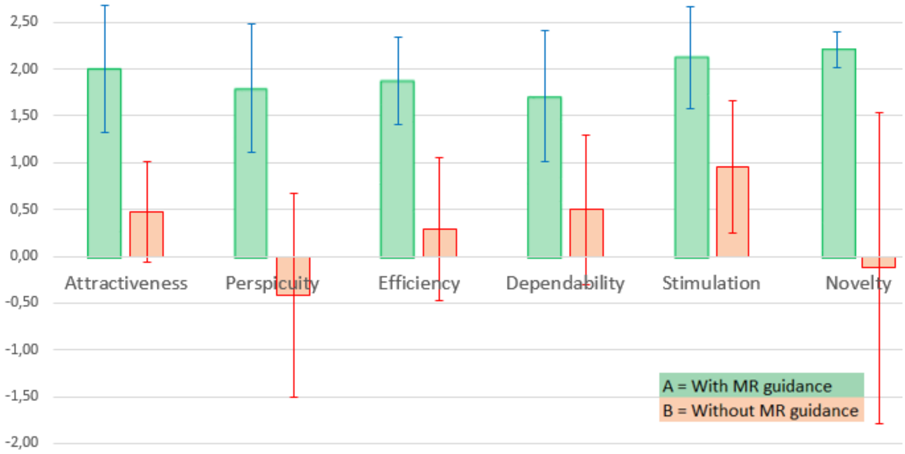
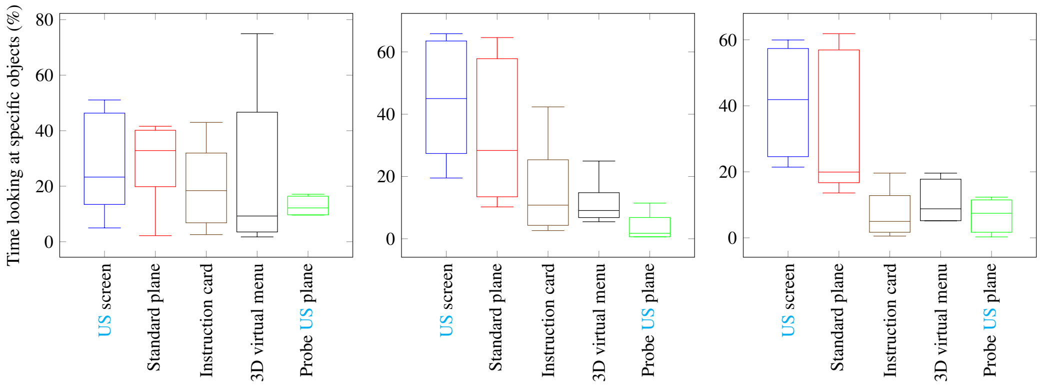


Figure 3.11: Product assessment result, represented by the comparison of scale means: the chart shows the scale means and corresponding 95% confidence intervals.



(**a**) Head standard plane (**b**) Abdomen standard plane (**c**) Femur standard plane

Figure 3.12: Visual attention profiles of study participants during standard plane navigation: Amount of time (in %) spent looking at specific game objects during navigation to the three standard planes, namely (**a**) head, (**b**) abdomen; and (**c**) femur.

guided standard plane navigation tasks are presented as box and whisker plots, showing the distribution of time spent looking at specific objects. Despite the small population size, some observations can be made: navigation to the head standard plane appears to require the least amount of time users had to look at the real US screen to find the location of the standard plane (Figure 3.12a). On the other hand, during navigation to the abdomen and femur standard planes, users spend more time looking at the US screen than looking at the actual 3D virtual standard plane (Figure 3.12b,c). With respect to the 3D virtual menu, the head standard plane shows the highest variation in time spent looking at it (Figure 3.12a). This could suggest that some users initially needed time to familiarise themselves with the menu interface, requiring less interaction in subsequent tasks. The smallest variation in attention to the instruction card occured during the abdomen standard plane task. This may imply that users found this plane easier to understand or identify, requiring less introductory guidance

Although the sample size is too small to draw statistically significant conclusions, these observations highlight the potential of eye gaze data to reveal meaningful user behaviour patterns. Such insights could contribute to refining future MR training systems and improving user experience. In particular, they may influence interface design choices, help identify cognitive bottlenecks during complex spatial tasks, or support adaptive feedback mechanisms—that is, systems that are able to respond dynamically to the user's behavior, such as offering 3D virtual real-time guidance or prompts when a user appears confused, distracted, or fixated on irrelevant areas.

3.5 Discussion

The current implementation of CAL-Tutor presents several technical limitations that influence both the usability and effectiveness of the system. These limitations affect multiple areas, including the accuracy of US probe tracking, the reliability of phantom model registration, and user interaction. Identifying and addressing these issues is critical to improving the training experience and developing future iterations of the system with improved performance and broader applicability.

HoloLensARToolkit provides a simple and convenient tracking method using the front-facing camera of the HoloLens 2 [166]. While this was sufficient to demonstrate the probe tracking, the accuracy of this single camera tracking could be improved. Tracking only works when the user is looking directly at the probe. Furthermore, the processing on the HoloLens leads to some latency and the probe must be moved relatively slowly to maintain tracking. Improved tracking emerged as a common suggestion made by the study participants.

One potential way to improve the range and accurcy of marker tracking is to use all HoloLens sensor cameras in research mode[9] which enables access to raw sensor streams that are typically unavailable in standard application mode. Using multiple cameras could enhance tracking range and robustness under challenging conditions such as occlusion, but it would require careful alignment of the cameras, which is challenging due to their differing view ranges.

In an initial attempt to investigate the use of external trackers for the assessment and comparison of vision-based tracking methods, the NDI Aurora electromagnetic tracker was incorporated into the data collection. However, no further investigation, and this remains an area for future work. While the Aurora could be used as the main tracking device, the convenience of visual tracking offers broader applicability and therefore remains one of the primary focuses of this thesis.

The current manual alignment of the obstetrics phantom model to its physical counterpart is labour-intensive, prone to human error and should be automated in future versions. Registration using larger ArUco markers and HoloLensARToolkit tracking gave acceptable results, but some inaccuracies remain. Many users also noted perceptual inaccuracies when using the HoloLens. Manual post-registration alignment of 3D virtual content may help overcome some of these issues by allowing users to adjust the model to their own satisfaction.

3.6 Conclusions and Future Work

This chapter presented an MR system that addresses an area that remains underexplored to date: AR-assisted training in obstetric sonography. CAL-Tutor aims to provide a sonography experience that is as close as possible to conventional fetal US training by allowing trainees to use conventional hardware such as the Voluson E10 US scanner and a SPACE FAN-ST phantom, while leveraging MR technology via a commercially available OST-HMD. CAL-Tutor is tailored to standard plane navigation tasks, in which novice users must navigate the US probe towards predefined targets located on the fetal anatomy. Initial feedback from six engineers showed that these novice users found that the MR guidance improved many aspects

---

[9] `https://learn.microsoft.com/en-us/windows/mixed-reality/develop/advanced-concepts/research-mode`, accessed on 26 May 2025.

of system interaction, such as efficiency, clarity and stimulation. This platform also records user motion data, provides interesting insights into user behaviour that could be further analysed, such as distinguishing novice from expert user motion.

Given its reliance on commercially and publicly available hardware and software frameworks, CAL-Tutor has the potential to support US education in regions with limited resources, where conventional training may not be readily accessible. Another important consideration is access to US scanners. While the Voluson E10 is commonly used in clinical practice, more affordable alternatives—such as portable wireless devices like the Clarius mobile scanner (Clarius Mobile Health Corp., Vancouver, Canada)—could be explored to increase feasibility in low-resource settings. Future adaptations of CAL-Tutor could be tailored specifically for integration with such portable US platforms to increase accessibility and ease of deployment.

While most research in the field of assistive technology for US standard plane acquisition focuses on DL-based methods for automated probe guidance during clinical procedures, CAL-Tutor offers an MR-based solution aimed at enhancing clinicians' spatial perception and hand–eye coordination during training. Nevertheless, the integration of DL networks to further support MR guidance toward standard planes could be explored in future iterations of CAL-Tutor. Live streaming of US imaging content onto the 3D virtual US plane that is placed at the ray-emitting part of the transducer is currently not supported. Instead, the virtual US plane displays a fixed sample US image. Although it is assumed that a live stream would render the US too small for users to recognise relevant details, future iterations of CAL-Tutor could explore this functionality as a potential enhancement. However, a potential limitation is that the US content may not remain continuously visible, depending on the orientation of the probe and how the US plane intersects with the 3D virtual baby model.

The technical limitations discussed in Section 3.5 should be addressed in future iterations of CAL-Tutor. In particular, the accuracy and robustness of US probe tracking proved critical to the system's overall usability and the 3D virtual overlay-guided probe navigation. These challenges led to a broader reconsideration of the tracking approach used in this thesis, ultimately leading to the exploration of markerless alternatives discussed in the following subsection.

3.6.1 Shift in Research Focus due to Challenges in Tracking Accuracy

The challenges related to marker-based tracking of the US probe, as discussed in Subsection 3.5, led to a shift in research focus toward a markerless tracking approach. The absence of physical markers, such as ArUco markers and other fiducial systems, simplifies the clinical setup by eliminating the need for additional hardware, but introduces new software-side challenges. While marker-based solutions usually benefit from available tracking software, markerless tracking alternatives rely on DL models that estimate the 3D pose of the object to be tracked in each video frame. When exploring markerless 3D pose estimation solutions for a hand-held tool, it also makes sense to add simultaneous 3D hand pose estimation, since the tool and the object are connected during a grasping motion. The CAL-Tutor application already records the 3D pose of the hand palm and wrist, since the HoloLens 2 comes with built-in markerless 3D hand pose estimation and tracking. This built-in tracking is based on depth camera that operates in high-frequency near-depth sensing mode, called AHAT (Articulated HAnd Tracking)[10], but also subject to the camera's limited field of view which limits the tracking range of the hand. In addition, grasping a hand-held leads to partial occlusion of the hand, which hinders the built-in hand tracking. Furthermore, HoloLens 2 hand tracking tends to struggle with small hand gestures and rapid movements. Additionally, limited on-device computational resources can cause delays in gesture detection and processing.

Limitations of the HoloLens 2's built-in hand tracking, along with the motivation to remain independent of its camera hardware, led to the exploration of solutions for markerless 3D hand pose estimation, alongside 3D ultrasound probe pose estimation. Therefore, markerless simultaneous 3D hand and tool pose estimation—based on DL models—was explored as a potential solution. Such models,

---

[10] https://learn.microsoft.com/en-us/windows/mixed-reality/develop/advanced-concepts/research-mode, accessed on 13 September 2024

on the other hand, require a sufficiently large amount of training data, in this case labeled images with the respective 3D pose of the object as ground truth. Creating such a dataset can be challenging. The next chapter examines the possibilities of egocentric 3D hand and probe pose estimation for this application, with a particular focus on the generation of a training dataset.

Chapter 4

HUP-3D: A 3D Multi-View Synthetic Dataset for Assisted-Egocentric Hand-Ultrasound-Probe Pose Estimation

4.1 Introduction and Background

This chapter presents a novel approach for markerless 3D hand and US probe pose estimation, developed to address the limitations of marker-based tracking methods in medical MR applications. The focus is on enhancing egocentric tracking performance, particularly in the context of obstetric US training, where accurate and real-time hand–tool interaction data is critical for effective user guidance and system feedback. The chapter begins by outlining the motivation for moving beyond traditional marker-based tracking systems, followed by background context and a description of the synthetic dataset generation process and evaluation pipeline.

The content of this chapter is based on the author's peer-reviewed publication in the proceedings of *MICCAI 2024* [55]. To support reproducibility, the HUP-3D dataset and accompanying codebase are openly available at `https://manuelbirlo.github.io/HUP-3D/`.

Marker-based tracking systems to determine the 3D position and orientation of the US probe, such as the AruCo code tracking used in the CAL-Tutor system (described in Chapter 3) suffer from key limitations: they rely on external hardware, require precise calibration, and are sensitive to occlusion and line-of-sight disruptions [240, 241]. These challenges motivated a shift (Section 3.6.1) toward markerless tracking in OST-HMD-assisted environments. This shift enables the use of video- and DL-based techniques to estimate hand and tool pose, broadening clinical applications beyond probe tracking. For example, marker-free, video-based tracking has been explored in posture tracking of surgical hands and tools [242], motion analysis of surgical instruments [243], and skill evaluation in robot-assisted surgery [244]. For obstetric US in particular, markerless methods offer significant advantages by eliminating the requirement to use additional tracking hardware. This reduces system complexity, lowers acquisition and maintenance costs, and simplifies deployment across different clinical environments, making such markerless solutions inherently more scalable and easier to integrate into existing workflows.

While existing systems for US probe guidance have relied on physical motion sensors [232], newer efforts have shifted toward egocentric vision-based tracking with MR headsets such as the Microsoft HoloLens 2 [245]. Estimating not only the probe pose but also the detailed 3D hand pose can unlock further potential innovations in clinical training:

1. **Refined training feedback:** Analysis of grip position, wrist alignment, and hand stability provides targeted feedback for improving probe control.

2. **Contextual understanding of motion:** Hand posture informs interpretation of probe motion patterns, enabling more effective movement guidance.

3. **Ergonomic optimisation:** Feedback on hand posture may help reduce fatigue or strain-related injuries.

4. **Simulation of real-world tasks:** Combined observation of hand and tool pose can better simulate authentic scanning scenarios and support sensorimotor skill development.

5. **Remote mentoring and guidance:** Visualising hand and probe movements may enhance teaching and collaboration, particularly in remote or expert-novice setups.

In US training, the lack of universally accepted competence metrics [233, 246] limits standardisation. Past attempts to derive objective measures from probe motion, such as Dromey et al.'s work using electromagnetic sensors [246], require attaching hardware to the probe. In contrast, purely image-based methods offer a marker-free pathway toward such measures.

4.1.1 Technical Foundations of Markerless Pose Estimation

Markerless 3D pose estimation refers to the process of estimating an object's 3D position and orientation directly from image data, without relying on physical markers such as fiducial tags or optical spheres. Instead, DL and computer vision techniques infer pose from RGB [247], depth [248], or RGB-D [249, 250] data. Most egocentric hand-tool pose estimation methods rely on RGB only [251, 240, 252, 241], especially in OST-HMD setups like the HoloLens 2, where RGB-D alignment is difficult due to different sensor fields of view (FOVs) and calibration needs [253].

The process of estimating the 3D coordinates of the hand and an interacting object typically involves a sequence of computational steps, as outlined below [51]:

1. **Detection of Region of interest**: Based on a single RGB image capturing a hand interacting with an object, the region within the image containing the hand-object interaction is identified and marked via a bounding box. Usually a CNN [254]-based approach is used to identify the region of interest [255].

2. **2D Keypoint Estimation**: The part of the input RGB image within the detected bounding box serves as an input for a CNN to predict the keypoints (2D coordinates) for the hand joints and object bounding box corners.

3. **Regression of 3D Hand and Object Pose**: An end-to-end differentiable neural network [254] is used to regress from the predicted 2D keypoints to the corresponding 3D coordinates.

4. **Refinement of 3D Hand and Object Pose**: By incorporating physical constraints between the hand and grasping object during hand-object interaction [251], the identified 3D hand and object pose gets further refined to ensure plausible grasp poses.

4.1.2 Synthetic Data Generation for Training Generative Pose Estimation Models

Such 3D pose estimation methods fall under supervised ML, meaning the entire pose estimation pipeline requires training on labeled RGB images. In this context, labeling involves providing each RGB training image with metadata that contains the ground-truth 3D poses of the hand and object. Ground-truth labeling can be achieved either through physical sensors attached to the hand and objects, capturing precise pose data, via manual annotation or through automated ML-based annotation techniques that infer 3D ground-truth directly from RGB images, as demonstrated by Hampali et al. [256]. Alternatively, synthetic RGB image generation methods can produce labeled datasets computationally [251, 240, 241, 257], avoiding the need for manual or sensor-based ground truth annotation. Since synthetic RGB images are generated from known 3D models of the hand and object, ground-truth annotations (3D poses) can be automatically and accurately generated without manual labeling. Such synthetic data generation often leverages computationally constructed RGB images depicting interactions between 3D mesh models of a hand, an object, lighting, and background. Blender[1] is a widely used tool for generating and rendering such hand-object interaction scenes [251, 240, 241, 257]. For hands, a widely used model in the research community is the MANO hand model [258], which follows the common convention of defining 21 keypoints consisting of joints and fingertips. To obtain 3D mesh models of the graspable objects a 3D scan of the real object is usually required. One commonly used method in the research community is structured light scanning, which reconstructs 3D geometry by projecting a known light pattern onto the object and analysing its deformation. For example, this method has been used to generate a 3D mesh model of a surgical

---

[1] `https://www.blender.org/`, accessed on 15 March 2025.

drill [257]. Alternatively, researchers in markerless 3D hand and object pose estimation often utilise pre-existing 3D mesh models instead of scanning real-world objects. Several publications in non-medical contexts [259, 260, 256] have relied on available 3D object models, such as mugs, bottles, and tennis balls, from the Yale-CMU-Berkeley Object and Model Set [261]. Another critical design factor in dataset creation is viewpoint diversity. Since applications involving OST-HMDs rely on egocentric perspectives, training data must reflect this viewpoint to ensure realism and effectiveness, while also covering a diverse range of angles to improve generalisability.

4.1.3 Challenges and Limitations of Markerless Pose Estimation

While markerless 3D hand and tool pose estimation offers several advantages over marker-based methods—such as a simplified setup or greater flexibility—it also presents a several technical challenges. These include issues related to occlusion, hand articulation, background clutter, data annotation and generation of plausible synthetic images for the training of DL pose prediction models. The following paragraphs outline these challenges in more detail, along with visual examples taken from recent literature.

Mutual occlusion. During close interactions, the hand and tool often obscure each other, complicating accurate 3D pose estimation [256, 262, 240, 252]. To illustrate this issue, Fig. 4.1 shows synthetic sample images of surgical hand-tool grasps exhibiting mutual occlusion, along with corresponding hand and tool pose estimations. While the pose predictions remain relatively accurate, the visual overlap between the hand and tool underscores the inherent challenge of achieving consistently reliable estimations in the presence of occlusion.

High degrees of freedom. Due to the fact that human hands have a high articulation with more than 20 degrees of freedom several parameters are required to model the complexity of plausible hand poses [264]. Small errors in estimating individual hand joints can significantly affect the accuracy and realism of the overall hand pose [265]. Fig. 4.2 illustrates this challenge of modeling high degrees of freedom in human hand articulation. All fingers of the hand must have joint angles that,

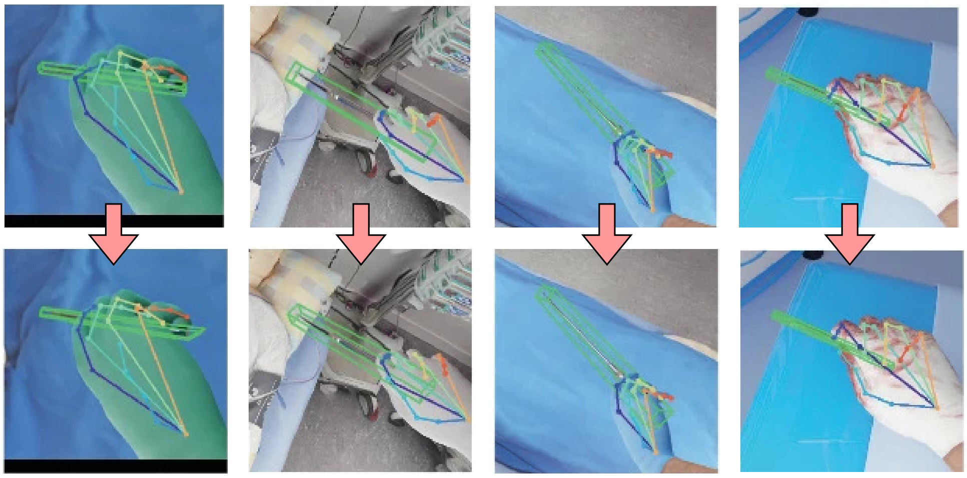


Figure 4.1: Selected qualitative results of hand and tool pose estimation on a synthetic surgical dataset. The figure uses 2D reprojections of hand joints and tool bounding boxes and illustrates mutual occlusion between the hand and tool. Top row: Synthetic input images with ground truth hand poses and tool bounding boxes. Bottom row: Reprojections of predicted 3D hand joints and object bounding boxes using the pose estimation method proposed by [263]. Figure adapted from Wang et al. [241].

together, form an anatomically plausible pose. Even a single joint in an implausible position can make the entire hand pose appear unrealistic.

Background clutter. In egocentric views images can contain irrelevant background information (objects, colors and textures) or visual noise (reflection, lightning variations etc.) that are not relevant to the hand-tool interaction itself, but may impact a proper identification of the hand and tool pose [262]. Such background clutter can make it more difficult to isolate the hand and tool from irrelevant surroundings within the image. In addition, clutter can distract or confuse ML models, which negatively affects their accuracy during training and inference. Additional preprocessing like segmentation may be required to distinguish the foreground including hand and tool from the background. Fig. 4.3 illustrates the challenge of background clutter in egocentric, visually crowded scenes.

Annotation challenges in real datasets. 3D annotations made by humans are difficult to obtain given the requirement of a large number of annotated images needed for model training. This labor-intense work makes it difficult to annotate a suf-

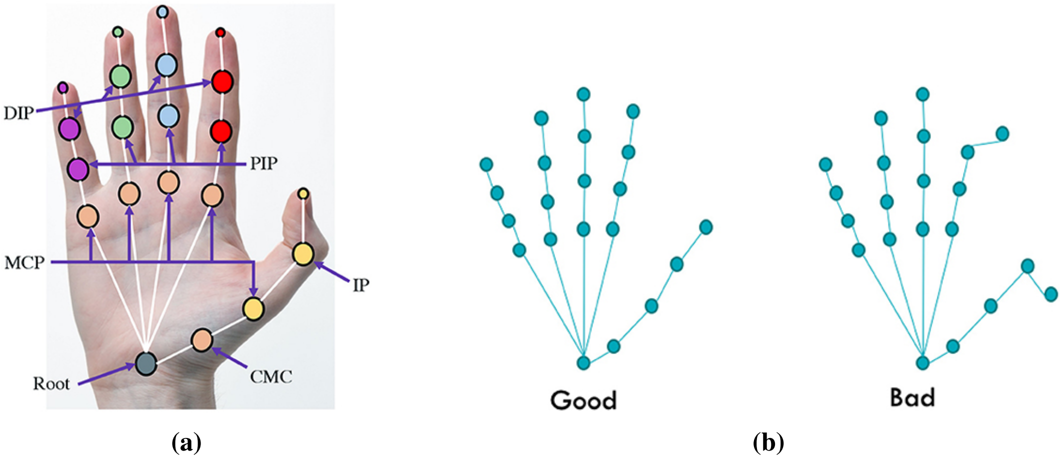


Figure 4.2: Illustration of hand articulation with high degrees of freedom. (a) Common definition of hand joints used for 3D hand pose estimation, with joint names: Metacarpophalangeal (MCP), Distal interphalangeal (DIP), Proximal interphalangeal (PIP), and Carpometacarpal (CMC). (b): Comparision of anatomically correct ("Good") and incorrect ("Bad") hand pose. Despite the anatomically incorrect pose on the right having all joints except two matching the original pose on the left side, only two joints that are at implausible angles make the entire hand pose implausible. Figures reused from Isaac et al. [265]

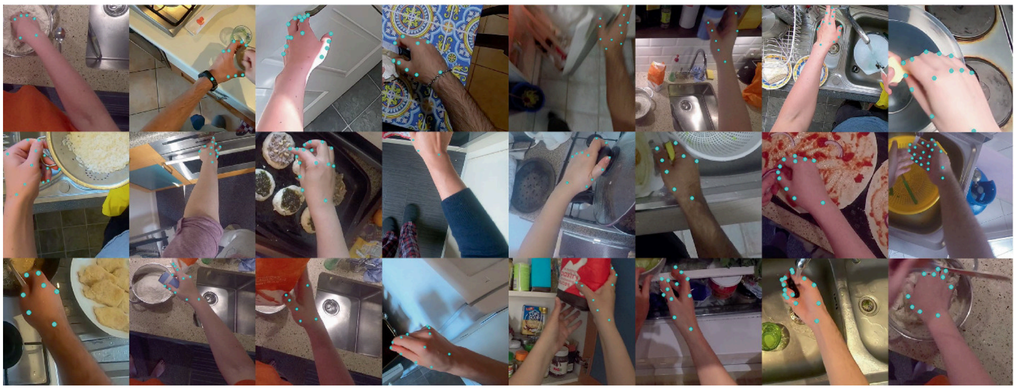


Figure 4.3: Egocentric images showing hand-object interactions in everyday kitchen scenarios from the EPIC-Kitchens dataset [266], illustrating visual background clutter. Blue dots indicate 2D hand joint annotations, which are part of the original image but not central to the purpose of this figure. Figure reused from Prakash et al. [267].

ficient amount of labeled images that would be required to allow the ML-model to generalise well to unseen images. Besides the scarcity of human annotations, precision of manual annotation is a limitation since humans cannot label the 3D ground truth of hand and tool perfectly based on a single image [263]. Instead of human annotation, physical sensors and optical markers are used to annotate real

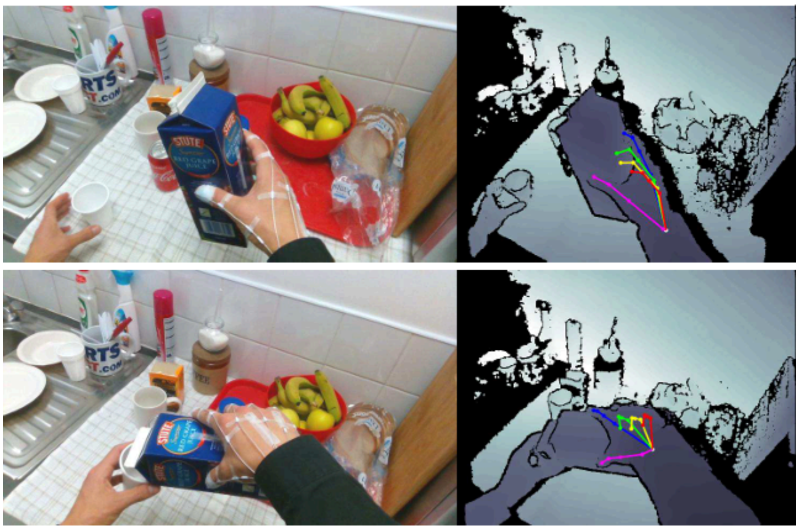


Figure 4.4: RGB-D video sample frames illustrating hand pose ground truth annotation using sensors. Left: RGB images with hand joint annotations, obtained via magnetic sensors strapped to the right hand. Right: Corresponding depth images showing the captured 3D hand pose. Figure adapted from Garcia-Hernando et al. [268]

data. Reflective markers can be attached to the tool to obtain its ground truth 3D pose [257]. To obtain the ground truth hand pose, electromagnetic sensors are used that are attached to each finger via wires. A sample scenario is shown in Fig. 4.4. While such physical data recording aids may provide a reliable ground truth they alter the images since markers and sensors are visible in the images and hence lead to an unwanted bias for ML-based pose prediction methods [259]. Another method to obtain ground truth is to record the hand and tool from multiple RGB-D cameras and generate a 3D point cloud. Ground truth annotation can then be obtained via a combination of manual and automated registration of 3D vertices of the tool and hand model [240]. Using multiple RGB-D cameras for capturing and annotating ground truth data introduces several challenges, including high equipment costs, complexity in camera calibration and synchronisation, missing data in scenarios of occlusion between hand and tool, labor-intensive and time-consuming manual annotation efforts, potential inaccuracies from algorithms aligning 3D models with point clouds, and scalability issues when creating large annotated datasets due to significant manual involvement.

Generation of plausible synthetic training data. Due to the aforementioned challenges of ground truth annotation in real training image data, recent efforts in the research community resulted in the generation of synthetic datasets [251, 240, 241, 257]. Despite the benefits of synthetic data over real data, such as implicit 3D model-based ground truth annotations, a major challenge is generating plausible hand grasp poses. Depending on the object to be grasped, plausible hand positions, orientations, and finger configurations vary, requiring suitable methods to generate realistic grasps. One approach that has been explored is the use of a software-based robotic grasping simulator to generate plausible grasps for everyday objects [251] and a battery-powered surgical drill [240]. Instead of a robotic gripper or robotic hand, the MANO hand model was used. However, its dexterity is limited in the simulator environment, as it follows a rigid robotic motion process. As a result, such simulation tools are constrained to graspable objects that require standard full-hand grasps at predefined positions, as is the case with a surgical drill or some everyday objects like a smartphone. Fig. 4.5 shows sample synthetic images of hand-object interactions involving everyday objects, where the grasp poses were generated using a robotic grasping simulator. While the grasps appear somewhat plausible, the hand poses are limited to robot-like, full-hand grasps and lack the full dexterity of a human hand.

Hand-held objects that allow for a greater variety of plausible grasp configurations pose a challenge for robotic grasping simulators, as will be explored in Chapter 4. One approach to generating plausible hand grasp poses is using generative ML models such as GrabNet [269] or Contact2Grasp [270], which take 3D models of the MANO hand and the object as input and predict plausible grasp poses based on automatically inferred object contact regions. Such models were also used to find plausible hand grasp poses for surgical tools like a scalpel [241]. The challenge, however, in using such generative models lies in the significant number of training iterations required with various graspable object models. One practical solution is to use a model that was pre-trained on everyday household objects [271, 272, 269], and to integrate it into a custom grasp rendering pipeline for specific objects and

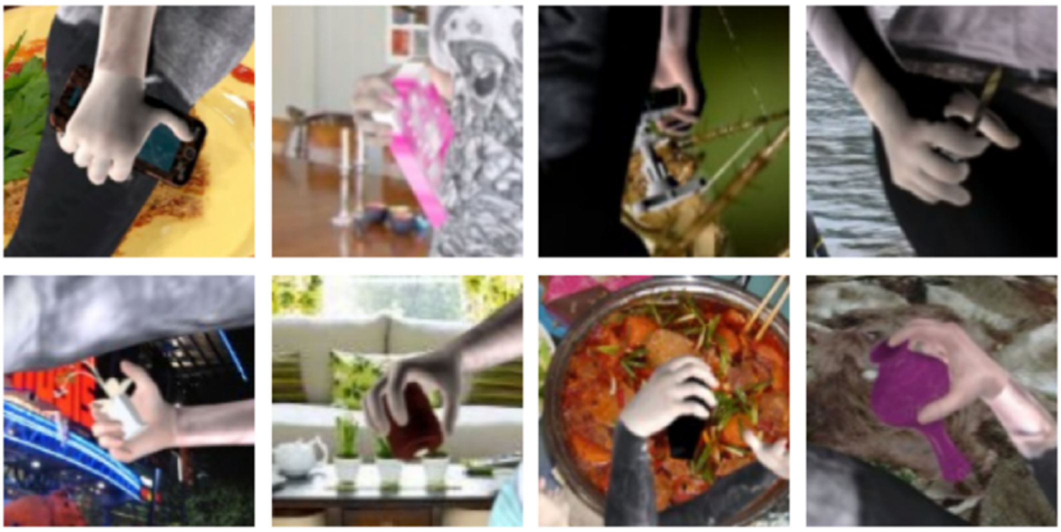


Figure 4.5: Sample images from a synthetic hand-object dataset, illustrating the challenge of generating plausible hand grasps. Figure adapted from Hasson et al. [251]

environments, such as a surgical setting in an operating theatre [241].

Another solution is to fine-tune such generative models with custom domain-specific annotated data in order to improve the model's capability to generalise to such newly introduced graspable objects. Such fine-tuning, however, could be challenging due to the need for additional labor-intensive contact region annotations between the MANO hand model and the newly introduced object models, complex hyperparameter tuning, and the risk of overfitting if too few new graspable objects are introduced.

These challenges and limitations naturally arise from the inherent complexity of advanced computer vision and ML methods, particularly when applied to dynamic, real-world tasks such as 3D hand and tool pose estimation. The variability in hand poses, occlusions during interactions, and the need for accurate annotations present significant hurdles for robust ML model development. Overcoming these challenges requires innovative approaches to data generation, model training, and evaluation.

4.1.4 Contributions

In response to the aforementioned challenges of markerless pose estimation, this chapter introduces a proposed solution—HUP-3D—which addresses issues such as viewpoint diversity and other key limitations inherent to markerless 3D hand and tool pose estimation. The contributions of this chapter and the associated HUP-3D project are as follows:

1. A scalable synthetic multi-modal (RGB-D, segmentation maps) image generation pipeline capable of producing diverse hand-US-probe grasp frames without requiring real-world data collection.

2. A novel sphere-based multi-view camera placement strategy that combines egocentric and non-egocentric perspectives for better generalisation.

3. The creation of the HUP-3D dataset[2]: a large, diverse, synthetic dataset consisting of over 31,000 images for joint 3D hand and tool pose estimation, featuring the Voluson$^{\mathrm{TM}}$ C1-5-D US probe, varied hand poses, textures, lighting, and camera angles.

4. Empirical demonstration of low 3D pose estimation errors (8.65 mm MPJPE) using a state-of-the-art model (HOPE-Net [273]) trained on the dataset.

A fundamental goal of these contributions was to develop methods that are highly reproducible and scalable, thereby providing added value to the research community and encouraging further research in markerless 3D hand and tool pose estimation. The code base associated with this chapter is designed to run on a Linux operating system. To support reproducibility across different host , the GitHub repositories of the image generation pipeline[3] include instructions for running the code within Docker[4] containers. The implementation of the methods underlying these contributions is described in detail in Section 4.2, followed by the experimental evaluation of the HUP-3D dataset in Section 4.3.

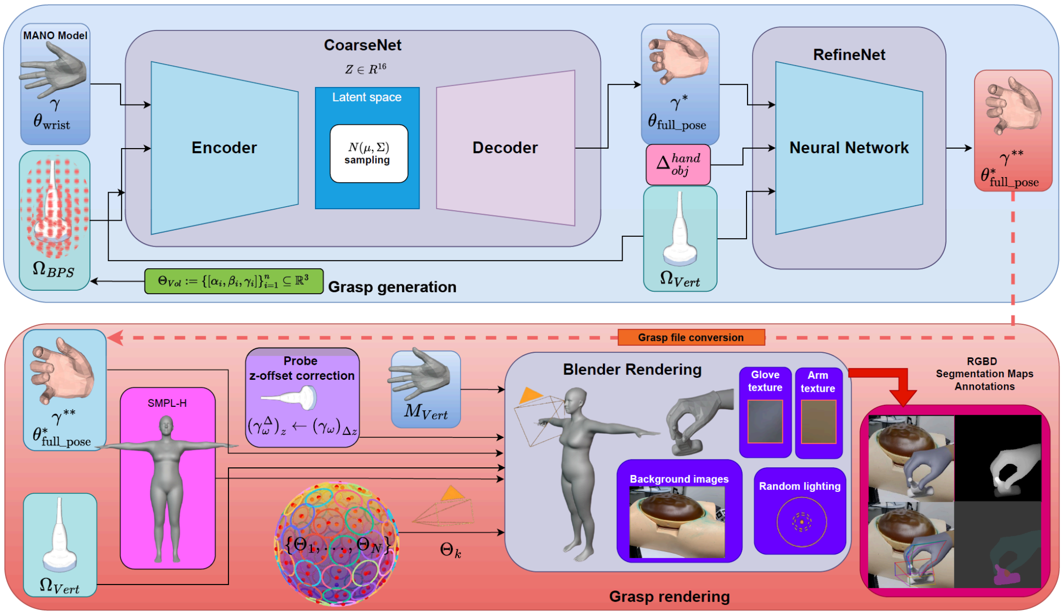


Figure 4.6: Grasp Generation and Rendering Pipeline Overview: In the grasp generation phase (upper part), the process starts with a MANO right hand model set to an initial pose $[\gamma, \theta_{wrist}]$ and a BPS-encoded [274] point cloud of the Voluson model Ω_{Vert}, oriented by predefined Euler angles Θ_{Vol}. These inputs are processed by CoarseNet to produce initial hand poses $[\gamma, \theta_{\text{full\_pose}}]$, which are then refined by RefineNet using vertex deltas for optimised hand-probe positioning, following the methodology from [241]. For grasp rendering (lower part), the final hand pose $[\gamma^{**}, \theta_{\text{full\_pose}}]$, along with model vertices and a SMPL-H body model, are imported into the rendering software. This stage adjusts probe positions for z-offset corrections, generates camera viewpoints from a spherical layout, and renders final images, ensuring only the right hand and arm are visible, with added textures and backgrounds, to produce RGB-D, segmentation maps, and annotations.

4.2 Method

This work targets potential applications in medical education, particularly within obstetric US, while remaining flexible enough to extend to other use cases. The synthetic image generation pipeline is divided into two main components: grasp generation and grasp rendering, which are described in the following subsections. A graphical overview of the complete pipeline is presented in Fig. 4.6. The final version of the grasp generation method employs a generative model to create re-

---

[2] All data in HUP-3D are synthetic. No human subjects were involved. Backgrounds were derived from anonymised public sources. The dataset poses no known ethical risks.

[3] The GitHub repositories can be found under the HUP-3D project page `https://manuelbirlo.github.io/HUP-3D/`, accessed on 10 June 2025.

[4] `https://www.docker.com/`, accessed on 10 June 2025.

alistic hand-tool interactions. However, the initial approach explored the use of a robotic grasp simulator, which produced suboptimal results in terms of grasp realism and variation. The limitations observed in this early attempt motivated the shift toward a generative modeling approach, as discussed in the next Subsection 4.2.1. This transition highlights the importance of choosing data generation methods that align closely with the complexity and variability of human hand interaction, particularly given the unique requirements for plausible grasp poses imposed by handheld objects.

4.2.1 Initial Grasp Generation Using GraspIt!—Evaluation and Limitations

The initial core idea of generating grasps was inspired by [251] and [240], namely using the robotic grasping simulator software GraspIt! [275], which requires a model of robotic hand or gripper and an object model to be grasped by the robot. Instead of a robotic end effector, the methods presented in [251] and [240] used a MANO hand model [276] along with tangible graspable object models. GraspIt! offers a principal component analysis (PCA)-based concept called "Eigengrasp" [277] to reduce computational complexity when computing robotic hand or gripper postures. An Eigengrasp can be considered as a low-dimensional representation of the higher-dimensional vector space of all possible hand movements. While this method may work well for simpler robot models with lower dexterity, it results in a restricted joint angle configuration space when used with the MANO hand model. This reduced dexterity implicitly reduces the variety of plausible hand grasps when using MANO. An initial attempt was made to generate plausible hand grasps using the GraspIt! simulator's graphical user interface. This was later replaced by a more suitable grasp generation approach based on a generative DL model, as detailed in Subsection 4.2.2.

For grasping objects that have fewer restrictions in terms of grasp position and grasp orientation, like a mobile phone or water bottle for example, an automated grasp finding solution using GraspIt!'s automated grasp finding options becomes feasible. The GraspIt! simulator offers several configurable automated grasp finding optimization algorithms that are based on a concept called "Eigengrasp planning" and were used in [251] for grasping everyday objects from various hand angles. The Eigengrasp planner needs to converge to grasp configurations that represent valid grasps that meet criteria such as grasp stability and force closure. An attempt was made to find valid hand grasps using a Voluson™ C1-5-D US probe model and the Eigengrasp planner. Several optimisation algorithms such as simulated annealing and gradient descent were used, but none of the selected algorithms converged. Figure 4.7 shows several sample hand configurations during a single Eigengrasp planner optimisation sequence in which the hand attempts to find valid grasps. It can be observed that the grasping attempts shown look unnatural due to the limited dexterity resulting from the Eigengrasp concept in combination with the shape of the Voluson. The shape of the grasp object and its underlying constraints on plausible grasp positions and orientations may disqualify the Eigengrasp planner from finding valid grasps. Instead, the GraspIt! simulator offers the option to position the hand and grasp object manually via translational and rotational adjustments in the graphical user interface. Once the hand and object are in a grasp-ready relative position, a so-called auto grasp can be executed that just closes the hand's fingers until they touch the grasp object. Hein et al. [240] generated synthetic images of a hand holding a surgical drill and used the auto-grasp option—rather than the Eigengrasp planner—because the task required grasps at specific orientations. The drill's symmetric, cylindrical handle made auto-grasp effective for generating plausible candidate grasp poses for manual selection.

Motivated by the results of manual grasp generation presented in [240], the same approach was applied using the Voluson™ C1-5-D US probe model. However, it had to be discarded due to several constraints.: Firstly, [240] considered grasp templates to be valid if they fulfilled Graspit!'s form-closure criterion[5]. The combination of MANO hand model and Voluson proved to be rather disadvantageous when it came to finding plausible form-closure grasps. Secondly, even the manual generation of plausible grasps—without the use of optimisation algo-

---

[5] `https://graspit-simulator.github.io/build/html/gfo.html`

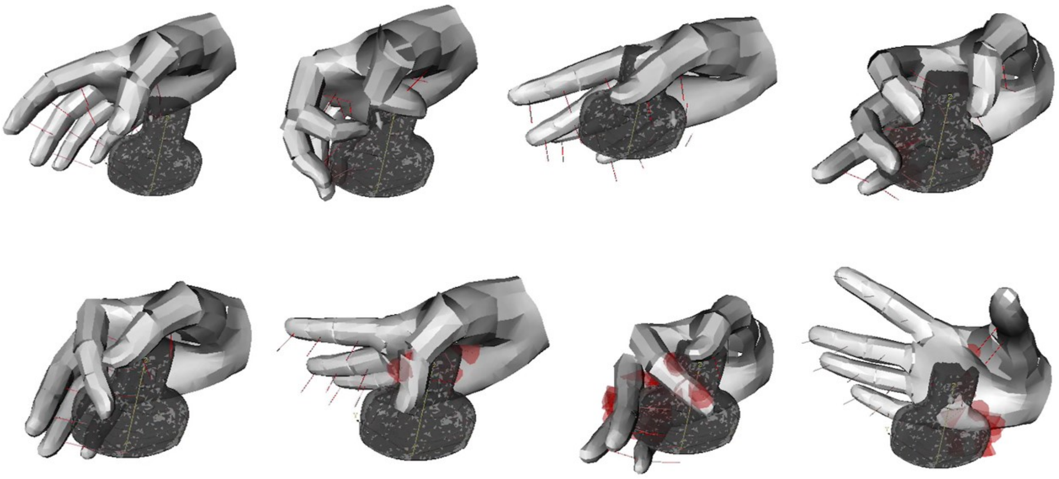


Figure 4.7: Sample hand grasp configurations of a GraspIt! Eigengrasp planner hand motion sequence using models of a MANO right hand and a Voluson™ C1-5-D US probe. The red lines shown perpendicular to each finger joint facilitate visual confirmation of the relative finger joint configuration. Red areas between the hand and the probe denote contact areas. Due to higher dexterity of MANO hand model and reducted dexterity of GraspIt!'s Eigengrasp concept, none of the simulator's available algorithms converges to stable and plausible grasps.

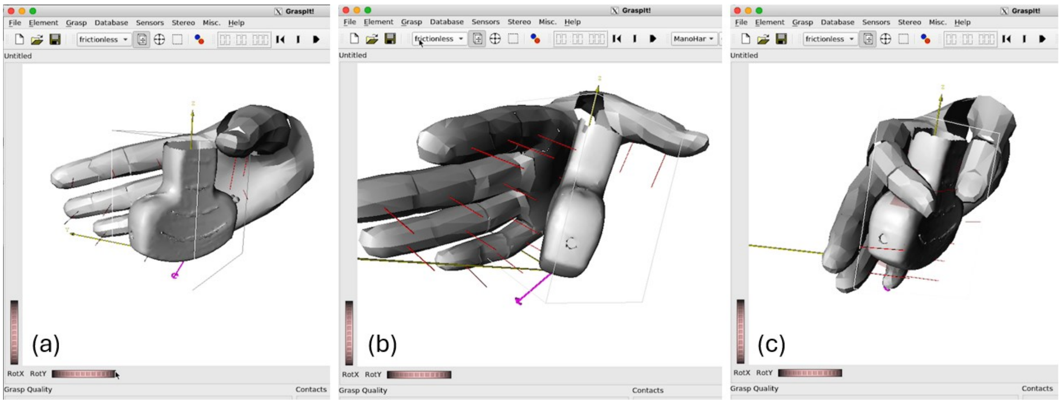


Figure 4.8: Sample manual auto grasp using GraspIt!, a MANO right hand model and a previously used 3D-reconstructed model of a Voluson™ C1-5-D US probe: (a)-(b): Initial manual hand pose before GraspIt! auto grasp ("Eigengrasp"), viewed from two different angles. (c): Suboptimal hand pose after GraspIt! auto grasp.

rithms—proved difficult after deactivating the form-closure criterion. This was primarily due to restricted dexterity when using the "auto grasp" option, which often resulted in unnatural grasp configurations. Figure 4.8 shows a manual sample grasp using the Voluson probe, before auto grasp (Fig. 4.8 (a)-(b) and after auto grasp (Fig. 4.8 (c))). The sample grasp shown in Fig. 4.8(c) illustrates a common issue observed across multiple manual grasp attempts: The grasp doesn't look plausible and also doesn't fulfill the form closure constraint. Therefore, an alternative grasp generation method was found by using a generative DL model, which is detailed in the following subsection 4.2.2.

4.2.2 Grasp Generation Using a Generative Model

To achieve automated annotation, a strategy focused on generating synthetic grasp images was adopted, avoiding the complexities associated with annotating real images. This approach allowed for a clear and manageable rendering workflow to be maintained. The underlying motivation in pursuing a purely synthetic image generation approach is to explore the possibility of creating a sufficiently large variety of training images to allow generalisability to real images for joint 3D hand and tool pose prediction. An initial feasibility study incorporated the use of a robotic grasping tool [275], which proved to be error-prone in this application and failed to produce a sufficiently large variety of plausible hand grasps due to limitations in hand dexterity. The generative model proposed in [269] for joint 3D grasp generation was subsequently adapted to a more clinical scenario. The proposed grasp generation process employs two sequential networks based on the MANO hand model [258]: an encoder-decoder network that generates initial coarse hand poses and a subsequent neural network dedicated to fine-tuning these poses, specifically enhancing accuracy in hand-tool interaction regions. The encoder, which samples from a normal distributed 16-dimensional latent space, requires encoded point cloud representations [274] of the probe model (Ω_{BPS}), together with the MANO right-hand model's initial translation $\gamma \in \mathbb{R}^3$ and hand wrist orientation $\theta_{wrist} \in \mathbb{R}^3$. Defined Euler angles Θ_{Vol} for probe meshes Ω_{BPS} were used for precise grasp pose control. Originally, the model described in [269] was trained with ordinary objects (like mugs, cameras etc.). However, its capability was extended to include the Voluson US probe. The decoder outputs an initial hand pose $[\gamma, \theta_{full\_pose}]$, which is subsequently refined through a neural network utilising the vertices of the probe model Ω_{Vert} and the vertex distances Δ_{obj}^{hand} between hand and probe. This refined pose, expressed as $\Psi := [\gamma^{**}, \theta_{full\_pose}]$, forms the foundation for our grasp render-

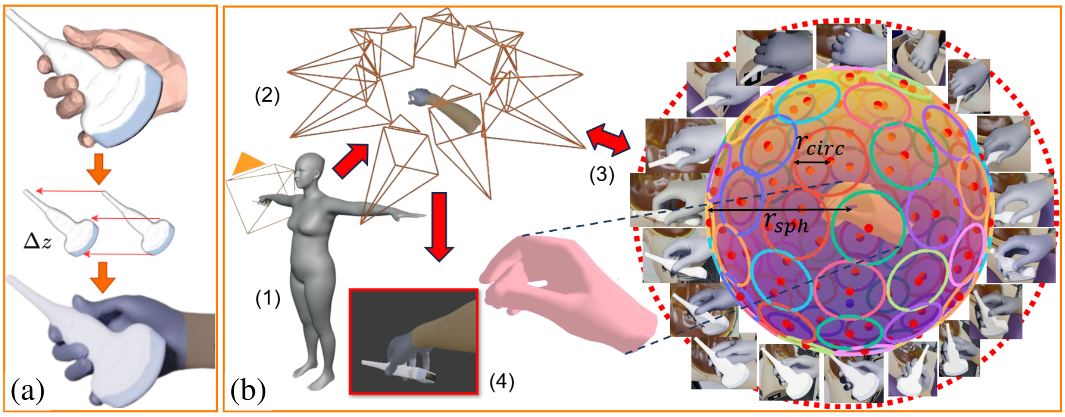


Figure 4.9: (a) Schematic grasp conversion from generative model to rendering software, including probe offset (Δz) correction. (b) Grasp rendering overview: (1) SMPL-H body model grasping the probe, showing egocentric and non-egocentric views. (2) Right arm and sphere-based camera orientations with remaining SMPL-H body parts hidden. (3) Camera angle sphere concept with views at various latitudes, centered on hand mesh; defines sphere (r_{sphr}) and circle (r_{circ}) radii. (4) Rendered hand-probe scene example from a sphere camera position.

ing approach detailed in Sec. 4.2.3.

4.2.3 Grasp Rendering

Using Blender, an open source 3D graphics software [278] for grasp rendering, as demonstrated in [251, 240], the rendering pipeline was tailored to accommodate the grasp poses Ψ produced by the generative model outlined in Sec. 4.2.2. Additionally, this rendering approach incorporates an SMPL-H body model [276], a MANO right hand model M_{Vert}, and the probe model's vertex data Ω_{Vert}. The grasp rendering pipeline can be seen in the lower part of Fig. 4.6. A short calibration step is required to reconcile the probe mesh used during grasp generation with the mesh used for rendering. Grasp synthesis assumes a probe coordinate frame with the origin located at the physical reference point (the probe's transducer face) and the probe axis aligned with $+z$. In the grasp renderer, however, the probe is placed at the world origin. Applying the synthesised hand pose Ψ directly can therefore produce a shift of the hand along the probe axis. To compensate for this axial shift, the offset Δz along the positive z-axis between the mesh origin and the reference point

---

[5] https://www.meshlab.net/

along $+z$ was calculated through polygon offset analysis (calculating the polygon differences in a simple Python script). Consequently, the probe's translation offset is adjusted by $\gamma_\omega^\Delta = \gamma_\omega + (0, 0, -\Delta z)$. Fig. 4.9(a) illustrates this correction. To enhance the diversity of camera perspectives, the approach was transitioned from a purely egocentric viewpoint strategy to the sphere-based methodology described in Camera view-angle sphere concept and illustrated in Figs. 4.6 and 4.9(b). This provides challenging examples such as mutual occlusions between hands and tools, improving the generalisability of the resulting pose estimation model.

For each grasp produced by the grasp generation module, a synthetic image is generated for every camera view angle $\Theta_k \in \{\Theta_1, \ldots, \Theta_N\}$, covering N positions around the sphere. Each image was rendered at two sphere radii, $r_{sph} \in \{0.5, 0.8\}m$, i.e., the egocentric camera distance (eye–to–hand distance) typical of US scanning. The rendering scene setup includes two shades of clinical gloves, randomised scene lighting, and eight backgrounds—(i) a lab with a SPACE FAN-ST US fetus model[6], (ii) consultation rooms, (iii) a white background, and (iv) real abdomens of pregnant women—sampled at random for each render. The rendering model outputs a comprehensive set of images for each grasp, including RGB-D and segmentation maps, as well as ground truth annotations. Sample frames from the HUP-3D dataset are shown in Fig. 4.10. A detailed overview of the parameters used in the grasp generation and rendering pipeline is provided in Section B.2 of Appendix B.

Camera view-angle sphere concept. The proposed methodology diverges from traditional egocentric viewpoints by implementing a sphere-based camera view setup to capture both egocentric and non-egocentric images, enhancing dataset diversity. This method, inspired by [279], involves distributing camera positions around a sphere, creating a varied perspective landscape around the right hand. The sphere is divided into horizontal segments, determined by latitude angles, to evenly distribute viewpoints. Specifically, the number of latitude segments N^ϕ and circles

---

[6] https://www.kyotokagaku.com/en/products_data/us-7_en/

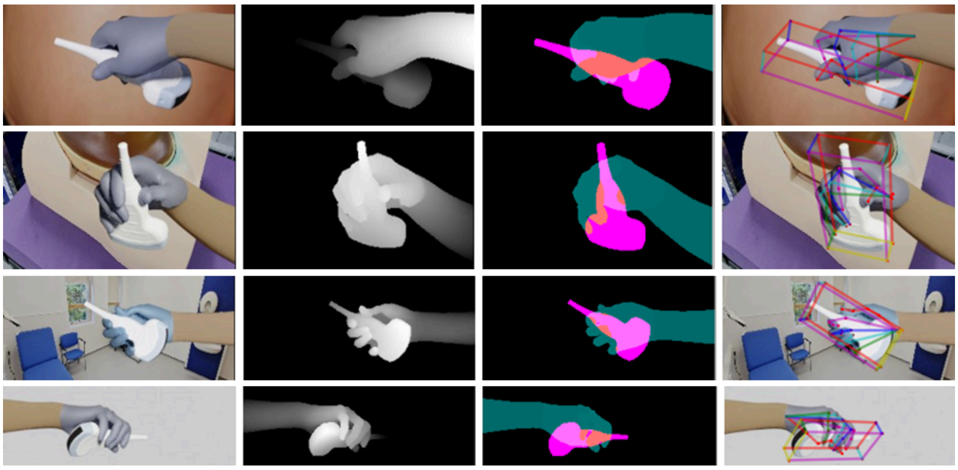


Figure 4.10: Sample frames from the HUP-3D dataset, grouped columnwise, from left to right: RGB, depth, segmentation map, and ground truth annotations.

per segment $N_{circ}^{(i)}$ are calculated to ensure comprehensive coverage:

$$N^{\phi} = \left\lfloor \frac{\pi}{2\arcsin\left(\frac{r_{circ}}{r_{sph}}\right)} \right\rfloor, \quad N_{circ}^{(i)} = \left\lfloor \frac{\pi r_{sph}\sin(\theta_i)}{r_{circ}} \right\rfloor, \; i \in \{1,2,\ldots,N^{\phi}\} \tag{4.1}$$

The division of the sphere into latitude floors was chosen to control camera placement and minimise frame redundancy, rather than to achieve perfect circle uniformity. The number of circles is determined in relation to the sphere's radius r_{sph} and the circle's radius r_{circ}. The circles are placed sinusoidally from the top to the bottom of the sphere. For each segment i and each circle j within, camera locations are defined by their spherical coordinates $(\theta_i, \phi_j^{(i)})$, ensuring a near-uniform spread of angles:

$$(\theta_i, \phi_j^{(i)}) \quad \text{with} \quad \phi_j^{(i)} = j \cdot \frac{2\pi}{N_{circ}^{(i)}}, \quad j \in \{0,1,\ldots,N_{circ}^{(i)}-1\} \tag{4.2}$$

This structured approach facilitates the generation of camera angles Θ_k, utilized in our subsequent rendering process. Figs. 4.6 and 4.9b visually demonstrate this concept, showcasing the strategic camera placement and the diverse grasp views it enables. A more detailed explanation of the camera sphere parameters is provided in Section B.1 of Appendix B.

Dataset	# frames	Source (Real/ Synth)	Viewpoints (Single/Multi/Ego)	Annotations	Modalities	Clinical (no. of tools)
HO-3D [256]	77.5k	Real	Single	automatic	RGB	-
ObMan [251]	153k	Synth	Multi	automatic	RGB-DS	-
ContactPose [280]	2.9M	Real	Multi	semi-automatic	RGB-D	-
Hein et al. [240]	10.5k	Synth	Ego	automatic	RGB-DS	1
POV-Surgery [241]	88k	Synth	Ego	automatic	RGB-DS	3
HUP-3D (ours)	31680	Synth	Multi	automatic	RGB-DS	1

Table 4.1: Dataset comparison: HUP-3D outstands as the first multi-view 3D hand-(clinical)object dataset.

Dataset comparison. Table 4.1 lists the top clinical and non-clinical datasets, together with their properties. HUP-3D is the largest multi-view dataset for clinical applications, presenting three modalities—RGB, depth, and segmentation (RGB-DS)—across multiple viewpoints. This makes it particularly well suited for training and evaluating pose estimation models that must generalise across varied perspectives and visual conditions. Only POV-Surgery [241] contains a higher total number (88k), but with less samples per tool (29k)[7] and just first-person view. In contrast, HUP-3D offers a broader range of perspectives and a more comprehensive tool-specific sample distribution. On a more general note, HUP-3D is the only publicly available synthetic dataset specifically tailored to obstetric sonography, featuring realistic right-hand grasps of an US probe. This unique clinical relevance, combined with the dataset's structured generation pipeline and automated annotations, makes HUP-3D a valuable resource for advancing markerless hand-tool interaction research in medical training and simulation contexts.

Only POV-Surgery [241] contains a higher number, but with less samples per tool (29k) and just first-person view. On a more general note, HUP-3D is the only publicly available synthetic dataset specifically tailored to obstetric sonography, featuring right-hand grasps of an US probe.

Noise characteristics. An additional aspect when comparing the datasets in Table 4.1 is whether image/depth noise was generated to increase realism. HO-3D [256] consists of real RGB images without any additional noise. However, the author's image capture and annotation pipeline used RGB-D cameras and a depth residual term to account for the noise in the captured depth maps during optimisation. Ob-

---

[7] 88k across three clinical tools, i.e., approximately 29k per tool.

Man [251], on which HUP-3D is based, is a fully synthetic dataset. The authors render RGB, depth and segmentation maps directly from Blender without introducing any additional noise into the image rendering pipeline. Hein et al. [240], also based on ObMan, is another fully synthetic dataset. However, unlike HUP-3D and ObMan, they randomise the egocentric camera pose by adding uniformly sampled noise to the hand and head positions of the 3D scene. ContactPose [280] consists of real RGB-D recordings and may therefore contain natural sensor noise; however, the authors do not mention noise characteristics. POV-Surgery [241] renders clean RGB, depth and segmentation maps without additional noise. Like ObMan, HUP-3D does not introduce additional noise during image rendering and attemps to increase image diversity through the aforementioned camera view-angle sphere concept and random variations in background image and lighting conditions.

4.3 Experiment: Pose Estimation Evaluation

To demonstrate the utility of the proposed HUP-3D dataset, a state-of-the-art DL model—originally designed for datasets such as HO-3D [256]—was deployed. As mentioned earlier, the dataset consists of 31,680 image sets from 11 realistic hand-object grasps. In a supervised learning setting, the data were split into seven grasps for training (20,160 images), two for validation, and two for testing (5,760 images). This will ensure the generalisability of the tested DL model. Table 4.2 lists the hyperparameters that were used to train HOPE-Net on HUP-3D.

Hyperparameter	**Value**
Batch size	64
Initial learning rate	1×10^{-3}
Learning rate decay step	10
Learning rate decay factor	0.95
Total training iterations	1000

Table 4.2: HOPE-Net training hyperparameters.

There have been extensive DL methods proposed for 3D hand-object pose estimation in the computer vision community. One of these competitive baselines is

HOPE-Net [273], originally tested on real data. HOPE-Net extends the capabilities of residual CNNs [281] with an adaptive Graph U-Net module [282]. This module manages to reduce the highly non-linear regression of the 3D hand and object coordinates. HOPE-NET's Adaptive Graph U-Net is based on the Graph U-Net [283] model, which considers images as special cases of graphs with nodes on pixel-based 2D grids. The underlying assumption of Graph U-Net is that traditional pooling and unpooling operations in U-Net-based encoder-decoder architectures are not well-suited for graph data. To address this, graph pooling and unpooling operations were proposed, enabling more natural processing of graph-structured data. Building on this concept, the Adaptive Graph U-Net in HOPE-NET refines these operations for the specific task of 3D joint hand-object pose estimation. It converts hand and object poses from 2D to 3D through novel and more robust graph convolution, pooling and unpooling. These improvements enable the HOPE-NET's estimator to effectively map RGB images to 3D world coordinates by exploiting the structured relationships between the hand joints and the object's boundary corners. To do this, the mean squared error loss is minimised during training on both 2D and 3D coordinates. The training procedure follows the same configuration as described in the original HOPE-Net publication [273].

Quantitative evaluation. Once the model was trained, the error in millimeters between the predicted joints and the ground truth was measured. In the test set, a total error of **8.65 mm** was achieved as Mean Per Joint Position Error (MPJPE), with 5.33 mm coming from the hand and 17.05 mm from the object. Although the value of 8.65 mm may appear large in absolute terms, it is competitive for monocular, markerless hand-tool pose estimation. Relative to the size of a hand or US probe, this corresponds to a small percentage. Most of the error comes from the probe (17.05 mm) rather than the hand, largely because the hand occludes parts of the tool, which makes accurate pose estimation more difficult for vision methods. The quantitative metric above aggregates errors over all hand joints and the probe bounding box. Although per-joint errors were not computed in this work, it is reasonable to expect higher errors in regions subject to occlusion or self-occlusion—consistent with the observation above for the probe. In particular, fingertips are frequently occluded during hand-probe grasps and are therefore likely to exhibit larger errors than proximal joints. This expectation aligns with prior analysis in depth-based 3D hand pose estimation, where fingertips showed higher per-joint errors and occluded joints yielded the largest errors [284]. The multi-view design of HUP-3D mitigates such cases by providing diverse viewpoints, but does not eliminate them entirely. The practical significance of these errors is application-dependent; quantifying them on real US hand-probe images would require a dedicated study and therefore lies outside the scope of this thesis. Under the same evaluation, the testing error of 8.65 mm is the lowest compared to other clinical data sets such as POV-Surgery [241] (14.35 mm) and Hein et al. [240] (17.02 mm) where even more advanced DL models were used. This supports the view that a multi-viewpoint dataset improves both hand and object localisation accuracy. Additionally, a simpler baseline composed solely of ResNet-50 [281] was tested, where the error was higher at 9.69 mm. Nevertheless, this still outperforms existing clinical datasets such as POV-Surgery and Hein et al. A comparison of MPJPE across these methods is presented in Table 4.3.

Table 4.3: Comparison of MPJPE in millimeters for hand, object, and total pose predictions on clinical egocentric datasets. The proposed method of using the HOPE-Net model on our HUP-3D dataset achieves the lowest error. The ResNet-50 baseline represents a simplified architecture, also trained on HUP-3D.

Dataset / Method	**MPJPE (mm)**		
	Hand	**Object**	**Total**
HUP-3D (HOPE-Net)	5.33	17.05	**8.65**
HUP-3D (ResNet-50)	–	–	9.69
POV-Surgery [241]	–	–	14.35
Hein et al. [240]	–	–	17.02

Qualitative evaluation. A subset of randomly selected frames from the HUP-3D test set was used for visual inspection of the model's predictions. A custom Python script (not publicly released) was developed to overlay predicted hand keypoints and bounding boxes around the US probe onto the corresponding RGB images. These were displayed alongside the ground truth annotations for direct comparison. Figure 4.11 illustrates examples of accurately predicted hand and probe poses, providing visual confirmation of the model's effectiveness.

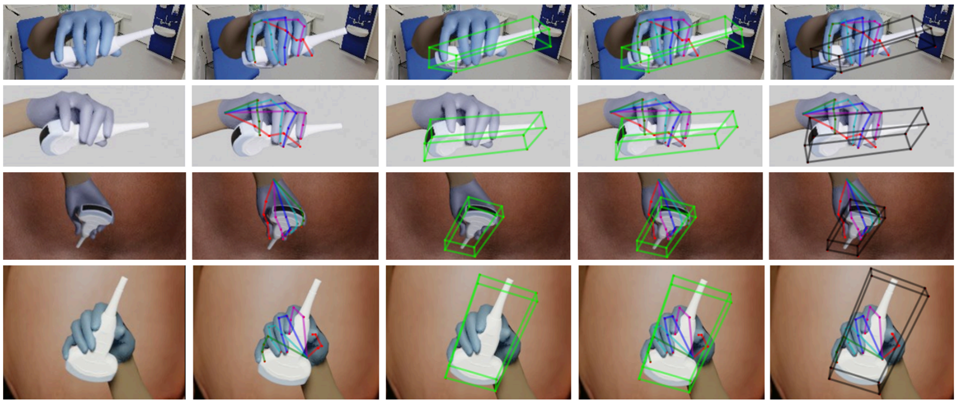


Figure 4.11: Qualitative results, shown with 4 test images from HUP-3D: image columns from left to right: RGB, predicted hand joints, predicted probe corners, predicted joints and corners, ground truth of joints and corners

4.4 Discussion

The results of the baseline evaluation in Section 4.3 demonstrate that the proposed HUP-3D dataset enables effective training of DL models for egocentric 3D hand-probe pose estimation. Reliable pose prediction was also achieved despite mutual hand–probe occlusion. The multi-view camera setup that includes non-egocentric synthetic images besides egocentric ones proved essential for achieving diversity in hand-probe grasps—crucial for training robust DL models.

In terms of reproducibility and extensibility, the proposed HUP-3D grasp generation and rendering pipeline—centered around the sphere-based camera view concept—offers an easy-to-use solution that can be scaled to automatically generate large volumes of synthetic data. This includes greater diversity in backgrounds, hand and glove textures, and lighting conditions. The method offers a practical balance between technical complexity and benefit to the research community, making it well-suited for broader adoption and future dataset extensions. Some of these extensions could further increase the variety of the dataset by introducing images of different hand shapes and left hands.

Synthetic datasets offer the flexibility to generate large volumes of plausible, ground-truth-labeled data, but they also come with limitations. The domain gap between synthetic and real-world images certainly impacts model generalisability. One primary aspect of generalisability is hand diversity, which is simulated through generative grasp synthesis, but the range of plausible grasp types remains constrained. This is largely because the underlying GrabNet model, used to generate synthetic hand poses, was trained on everyday handheld objects rather than on US probes. Moreover, GrabNet may not support full power grasps due to its hand contact region formulation, which tends to favour grasps characterised by fingertip and distal/middle phalange contact, rather than full-palm engagement typically seen in power grasps. Another limitation of GrabNet is its lack of directional controllability. During inference, the model takes only the wrist translation and rotation as input, resulting in generated grasps with random hand orientations [285]. Furthermore, GrabNet does not account for full-body constraints, such as those imposed by surrounding individuals or environmental context [286, 285]. As a result, it can generate hand–probe grasps with implausible orientations—such as reaching across or beneath the US probe—scenarios that are physically unfeasible due to the presence of the pregnant patient during a fetal US scan or because they involve contact with the transducer area, which is not a typical grasp region. Future iterations of HUP-3D should explore more advanced grasp generation methods.

Dataset extension and multimodal evaluation. Another important factor for improving generalisability is greater variation in hand grasps using different tools. In its current form, the HUP-3D dataset includes only a single US probe model, which limits variation in tool geometry. Including additional probes would help create a more diverse training set and improve model generalisability to unseen scenarios.

In addition, the experiments in Section 4.3 rely solely on a single modality (RGB), without fully leveraging the multi-modal nature of HUP-3D. Incorporating a second modality is expected to enhance both the generalisability and accuracy of the pose prediction model. Therefore, a dual-modality evaluation using both RGB and depth data is envisioned. These two directions—a dataset extension incorporating a second US probe and a comparative single- and dual-modality evaluation—are explored in the next chapter, Chapter 5.

4.5 Conclusion and Future Work

HUP-3D has been introduced as a pioneering 3D hand–object multi-view dataset tailored for modeling hand–US probe interactions in obstetric scenarios. It is designed to support research in clinical movement analysis using egocentric camera views and MR applications. The dataset generation pipeline combines a versatile grasp synthesis model with an automated rendering process, demonstrating the benefits of a multi-view camera sphere setup. A baseline model evaluation confirmed its effectiveness, even in the presence of significant hand–probe occlusions. While the presented dataset is fully synthetic, future iterations may incorporate real images to increase image diversity. Annotating such real images could be achieved via automatic ground truth generation methods, such as the approach presented by Hampali et al. [256], thereby avoiding labor-intensive and error-prone manual annotations using markers, additional cameras, or sensors.

The GrabNet-based grasp synthesis method produced plausible grasp poses, but also revealed limitations regarding grasp orientation, position, and contact surfaces, as discussed in Section 4.4. Using an US probe as a graspable object introduces additional constraints that are not captured by models pre-trained on everyday handheld objects. Future work could explore more advanced grasp generation approaches that better reflect the unique characteristics of clinical tools like US probes.

The current grasp rendering setup relies on a limited set of background images, which restricts the visual diversity and realism of the generated scenes. To enhance realism and better reflect clinical environments, 3D-reconstructed US room settings could be incorporated into the Blender scene, as demonstrated in [241]. This would require additional hardware to capture the room from multiple perspectives and software capable of reconstructing a 3D room model suitable for import into Blender. An added benefit of this approach would be the ability to include anatomical models—such as a baby and a maternal uterus—placed on a table within the scene. The SMPL-H body model could then be positioned in a physically plausible location relative to these objects. As a result, rendered frames would feature more realistic backgrounds and spatial relationships between the hand, probe, and clinical environment.

Another option to increase the realism of the synthetic images and improve the efficienncy of the rendering pipeline is to use BlenderProc2 [287], a Blender-based framework for photorealistic image generation. BlenderProc2 is specifically tailored for creating synthetic datasets used for training DL-based pose prediction models, and includes built-in functionality for generating color, depth, surface normal, and semantic segmentation images. Future iterations of HUP-3D could explore the potential benefits of integrating this software.

The current pose prediction model operates on single time-stamped images representing final grasp configurations. However, incorporating temporal modeling of grasp sequences could enable more realistic interaction dynamics and improve prediction robustness. This would require capturing and learning from temporal hand–object interaction data. Although such a direction lies beyond the scope of this thesis, a conceptual framework will be discussed in Chapter 6.

As noted in Section 4.4, the HUP-3D dataset is currently limited in terms of grasp diversity, as only a single US probe model was used. Additionally, the baseline evaluation relies solely on RGB input, despite the availability of multi-modal data. Exploring the added value of combining RGB and depth modalities may improve generalisability and pose prediction accuracy. Both extensions—a multi-probe dataset and a dual-modality evaluation—are explored in Chapter 5.

Chapter 5

HUP-3D-v2: Dataset Extension and Multi-Modal Evaluation for Hand–Probe Pose Estimation

5.1 Introduction

This chapter continues the work presented in Chapter 4 and addresses the third key objective of this thesis, as outlined in Section 1.3 of Chapter 1. It extends the HUP-3D dataset by incorporating a second US probe model and introducing multi-modal data—specifically RGB and depth images—for egocentric hand–probe pose estimation. Chapter 4 focused on the development of a synthetic data generation pipeline, the creation of the HUP-3D dataset and a single-modality (RGB) evaluation of a state-of-the-art 3D pose prediction model trained on HUP-3D. While the results demonstrated the feasibility and accuracy of DL models on HUP-3D, they were limited in two aspects: First, the dataset included images from only a single US probe, restricting generalisability; and second, the model evaluation was based solely on single-modality (RGB) input.

This chapter builds upon that foundation by addressing three open questions: how portal US probes affect the synthesis of plausible hand-probe grasps; whether multi-modal (RGB-D) data can improve pose estimation accuracy; and how increased data diversity impacts model generalisability. Although the work presented here is not part of a separate peer-reviewed publication, it represents a direct extension of the original HUP-3D pipeline. The updated dataset and additional evaluation tools are included as part of the open-source HUP-3D project page at `https://manuelbirlo.github.io/HUP-3D/` to support future research and reproducibility.

5.1.1 Context and Technological Motivation

The HUP-3D dataset presented in the previous chapter 4 consisted of images of hand grasps using the Voluson™ C1-5-D US probe, which is used in conventional obstetric sonography. Voluson probes are connected to an US imaging machine via a cable. During an obstetrics US scan such an imaging machine is usually placed away from the patient, and sonographers must look away from the patient in order to observe the live US imaging video while scanning the patient. In other medical applications of US imaging, practitioners face similar situations with the US screen being positioned away from the patient. This requires sonographers to develop refined hand-eye coordination skills [231].

Studies have shown that in other medical US scanning applications, such as vascular access and point-of-care US, there are differences in gaze patterns between novice and expert sonographers. Chen et al. examined gaze patterns during US-guided Internal Jugular Central Venous Catheterisation and found that novices spent significantly more time looking at the probe and needle, whereas more experienced clinicians focused primarily on the imaging screen. Similar findings were observed by Chan et al. in their assessment of gaze patterns during point-of-care US training [288]. Following these findings, it can be hypothesized that in obstetric sonography, more experienced sonographers would also spend significantly less time looking at the US probe. Consequently, MR training applications, such as the CAL-Tutor system presented in chapter 3, may provide more substantial benefits for novice users, who rely more heavily on visual feedback from the probe. However, if a user is wearing an OST-HMD and is not looking at the US probe, the egocentric markerless 3D hand and tool pose estimation method presented in chapter 4 becomes less effective, as hand and probe tracking relies on direct line-of-sight.

Portable US scanners may offer a solution in this regard.

Recent developments in US technology led to portable handheld scanners that can be used in remote point-of-care settings where conventional, more expensive and much larger US machines are not available. Live imaging can be done via mobile devices such as smartphones and tablets, which can be placed closer to the relevant patient anatomy. This proximity between patient and imaging screen has the potential to reduce the sonographers cognitive workload by reducing the demands of hand-eye coordination. One such portable device, and a popular choice among healthcare professionals, is the Clarius[1] US scanner, which is available in several versions. Portable US scanners offer dedicated software that runs on the mobile device so that obstetric biometry can be performed. Rittenhouse et al. analysed the accuracy of such portable US machines — the Butterfly iQ[2] and the Clarius C3 — via assessment of fetal biometry and estimated gestational age; in comparision with a conventional US machine [289]. In their study, experienced sonographers assessed both fetal biometry and estimated gestational age accross several patients using portable and conventional scanners. The results confirmed good accuracy for both types of obstetric assessments. However, the portable scanners showed increasing errors in accuracy with advancing gestational age.

Focusing on an obstetric-centered use of portable US scanners, recent research highlights the potential advantages of such devices in resource-constrained environments, including cost-effectiveness, adaptability, and usability [290]. These benefits make handheld scanners particularly suitable for deployment in areas, where access to conventional US systems and trained clinicians may be limited. As such, portable US scanners present a promising opportunity to expand prenatal care and diagnostic access globally.

Contribution of this chapter. The contribution of this chapter is twofold:

1. With the potential benefits of portable US scanners in mind, as discussed in Section 5.1.1, this chapter introduces an extension of the HUP-3D dataset. The updated version, named HUP-3D-v2, consists of the 31680 frames of

---

[1] (Clarius Mobile Health Corp., Vancouver, Canada)

[2] Butterfly Network, Inc., Guilford, CT, USA

the HUP-3D dataset (Chapter 4), but adds another 31680 frames of synthetic hand–probe grasp images generated using a 3D model of the Clarius C3 scanner. The goal is to provide a second dataset—twice in size to the original HUP-3D—that enables pose prediction models to generalise more effectively to unseen input data. A standalone dataset consisting of 31680 Clarius probe grasp images—called HUP-3D-Clarius[3]—is available as well. Both datasets, HUP-3D-Clarius and HUP-3D-v2, are available via the HUP-3D project page. In addition, the standalone HUP-3D-Clarius dataset is available for download here: HUP-3D-Clarius dataset[4]

2. A multi-modality evaluation of the extended HUP-3D-v2 dataset was conducted to investigate the effect of incorporating depth data alongside RGB input, in combination with increased dataset diversity. The corresponding architecture is described in the paragraph on multi-modality. In addition, a single-modality evaluation was carried out to assess the generalisability of the HOPE-Net model on the extended dataset.

Section 5.2 details the methodology behind the HUP-3D dataset extension, including the process of generating synthetic hand-probe interaction images using the Clarius 3D model. It also outlines the evaluation pipeline for both multi-modality and single-modality experiments. Together, these evaluations offer insights into how increased data diversity and modality combinations can influence model performance in markerless pose estimation tasks within the specific context of obstetric sonography.

5.2 Method

The grasp generation and rendering methods used for HUP-3D-Clarius are nearly identical to those described in Chapter 4, Sections 4.2.2 and 4.2.3. Only two differences apply: First, a 3D model of a Clarius C3 probe was used instead of the

---

[3] All data in HUP-3D are synthetic. No human subjects were involved. Backgrounds were derived from anonymised public sources. The dataset poses no known ethical risks.

[4] https://drive.google.com/file/d/1g5AUXEuxpSAseGebkp0xUo_m4DKV01EN/view?usp=sharing

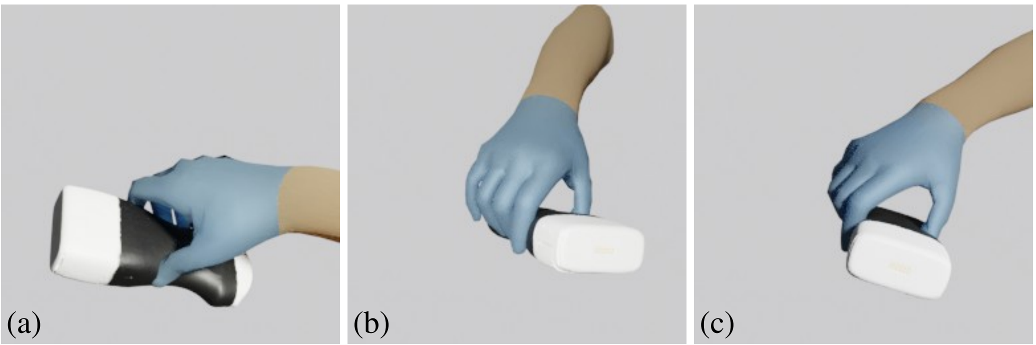


Figure 5.1: Interpenetration of Clarius probe and fingers, shown on a single grasp from 3 different camera view angles: (a) interpenetration of probe and thumb, (b) interpenetration of probe and index finger, (c) interpenetration of probe and both thumb and index finger.

Voluson C1-5-D probe. Second, the software-based offset correction described in Section 4.2.3 and shown in Figure 4.9 (a) was not applied to the Clarius model. Instead, the .obj[5] file of the Clarius probe was manually corrected using MeshLab by aligning the model to the origin. As a result, no additional offset correction was required in the rendering pipeline.

A total of 11 new Clarius grasp sequences, each consisting of 2880 frames, were generated—resulting in a dataset of 31,680 frames referred to as HUP-3D-Clarius. These frames were combined with the existing 31,680 frames from the original HUP-3D dataset (based on the Voluson probe) to create an extended dataset named HUP-3D-v2, with a total of 63,360 frames. The aim of the HUP-3D-v2 dataset enhancement is to increase variability by incorporating grasp interactions with a second US probe. This added diversity is expected to improve the generalisability of DL models for joint 3D pose estimation of the hand and ultrasound probe. In parallel, HUP-3D-Clarius is provided as a standalone dataset for applications focused specifically on Clarius-based probe tracking and pose prediction.

Limitations of the grasp generation method. Although the grasp generation method for HUP-3D-Clarius is largely the same as that used for HUP-3D, a limitation was observed when generating grasps with the larger Clarius probe model.

---

[5] An .obj file is a standard 3D model file format that includes geometric data such as vertices, texture coordinates, normals, and polygon faces.

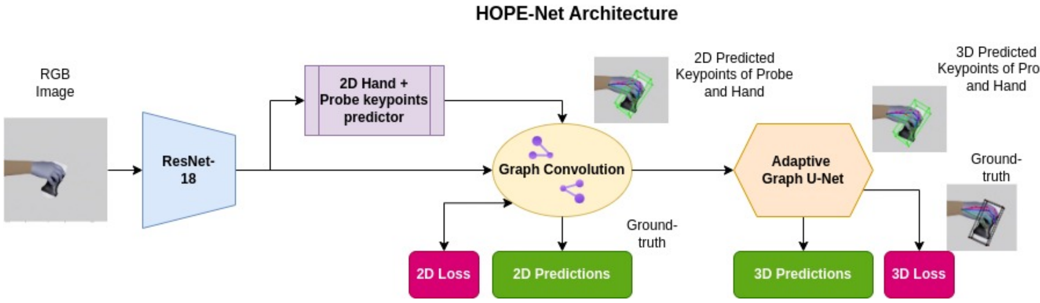


Figure 5.2: HOPE-Net architecture for single-modality input using RGB images only. The diagram is illustrated with a sample synthetic RGB image from the extended dataset as input.

GrabNet [269], the generative model used, was pre-trained on everyday handheld objects (binoculars, mugs, and toothpaste, etc.) and does not generalise perfectly to larger tools like the Clarius. Specifically, the increased width of the probe forces the hand model to open wider, occasionally resulting in minor interpenetration between the fingers (particularly the thumb and index finger) and the probe geometry. Figure 5.1 illustrates this issue in RGB frames from the HUP-3D-Clarius dataset, shown from three different camera angles. GrabNet already includes penetration penalties and contact refinement, which limits the impact of such artefacts, but residual collisions are a known limitation. In the HUP-3D pipeline, the impact of penetrations was further reduced by using simple offsets—either via software-based correction in the grasp renderer (Chapter 4, Section 4.2.3, Fig. 4.9a) or by aligning the probe model to the origin (this chapter). Since the level of interpenetration remains visually minor and does not significantly affect the perceived realism or structure of the grasps, retraining the generative model to eliminate the interpenetration was considered disproportionate in the context of this research due to substantial time and effort required. Instead, this interpenetration was accepted and the generated synthetic Clarius hand grasp images were used to test the utility of this newly acquired dataset.

5.2.1 Application of HOPE-Net for 3D Pose Estimation

To evaluate the extended dataset, the state-of-the-art DL model HOPE-Net [51] was employed for 3D hand and US probe pose estimation. Prior to the evaluation, the HOPE-Net architecture for the single-modality (RGB) case is briefly described. This is the same architecture that was previously used to evaluate HOPE-Net's performance on the HUP-3D dataset, as detailed in Section 4.3 of Chapter 4. Additionally, the subsequent paragraph on multi-modality (RGB-D) architecture outlines the extended architecture used for evaluating multi-modal input.

Single-modality configuration. Fig. 5.2 shows the core components of the HOPE-Net architecture, illustrated on a sample synthetic RGB image from the extended dataset. The RGB image is fed into a residual neural network, ResNet-18, that acts as an image encoder and extracts features from the input image, a high-dimensional (2048D) feature map, also called embedding. This feature map—a high-dimensional vector, encodes the image's spatial and semantic information—is then passed to the 2D hand and probe keypoints predictor, which regresses 2D keypoints (hand joints and corners of the tight bounding box surrounding the probe) using a fully connected layer, supervised by a 2D loss. The fully connected layer generates 29 keypoints (21 hand joints and 8 probe bounding box corners) which serve as input to a graph convolution layer. Simultaneously, the 2048-dimensional ResNet-18 feature embedding is also fed into a 3-layer graph convolution layer, providing contextual information to support spatial reasoning over the keypoints. More precisely, the graph convolution layer represents the hand and probe keypoints as nodes in a graph, enabling the model to infer spatial relationships between them and refine their positions based on structural context. Consequently, the graph convolutional layer outputs refined versions of the initially predicted 2D hand and probe keypoints, enhanced by the relational context encoded in the graph structure. The refined 2D keypoint features from the graph convolutional layer—representing node features for each of the 21 hand joints and 8 object corner keypoints—are then passed into the Adaptive Graph U-Net, an encoder-decoder architecture designed to capture spatial relationships using a series of graph pooling and unpooling layers. The network is "Adaptive" because pooling is input-dependent: at each pooling stage it learns scores for all nodes and keeps the most informative ones, then reconstructs the graph connectivity over the retained nodes. In each transition from one pooling layer to the next within the Adaptive Graph U-Net, the number of nodes is halved, retaining only those nodes that capture the most salient information about both the hand and the object's bounding box. To preserve spatial information lost during pooling, each pooling layer has a skip connection to its corresponding unpooling layer. A skip connection directly transfers feature representations from the pooling (encoder) layers to the unpooling (decoder) layers, enabling the network to recover spatial details that may be lost during downsampling. In the unpooling sequence—the decoder part of the network—the downsampled hand and object bounding box features are progressively reconstructed in each unpooling layer, where each layer doubles the number of features. In this sense, the decoder acts as the generative component of the Adaptive Graph U-Net, creating the 3D hand and object pose predictions. More broadly, it learns to map the structured 2D keypoint representation to 3D keypoint predictions through a hierarchical graph learning process. The 3D predictions generated by the Adaptive Graph U-Net are compared against the 3D ground-truth keypoints—provided by the metadata of the synthetic input image—to compute the final 3D loss.

Justification for multi-modality configuration. Prior work in the area of RGB-D-based multi-modal DL for 3D hand pose estimation has shown that multi-modality can increase model generalisation and contribute to model performance. In [291], a dual-modality network was proposed using synthetic RGB and synthetic depth images and a novel cross-modal keypoint fusion concept. This method outperformed state-of-the-art models on four datasets. [292] addressed generalisation issues related to a model's domain gap that arise when working with synthetic rather than real-world data and proposed a dual-modality network for effective pre-training of the model. Experiments showed that their proposed method improves 3D hand pose estimation on benchmarks. [293] presented a RGB-D-based hand pose estimation concept that leverages RGB and depth information in separate steps: First, a network is used for hand detection and hand feature regression. Second, additional depth map information is used in an energy minimisation framework for accurate hand pose and shape estimation. Baseline comparison experiments demonstrated

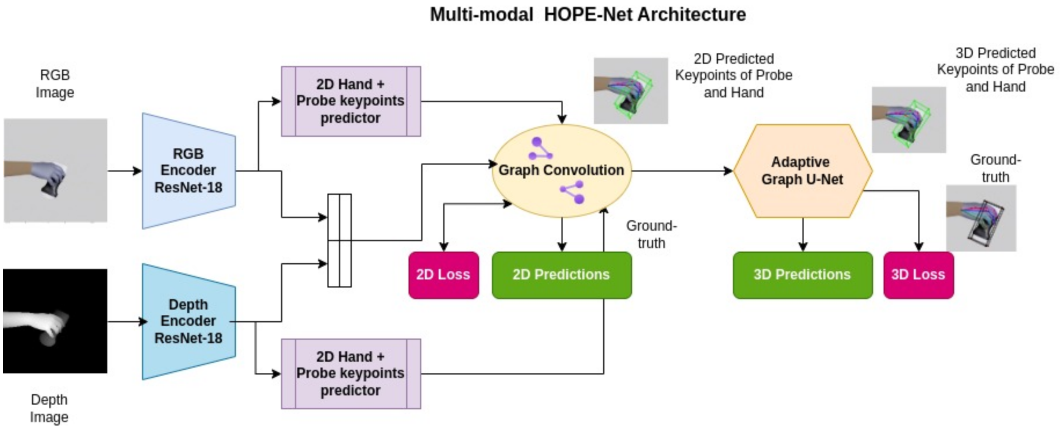

Multi-modal HOPE-Net Architecture
RGB Image
RGB Encoder ResNet-18
2D Hand + Probe keypoints predictor
Depth Encoder ResNet-18
2D Hand + Probe keypoints predictor
Depth Image
Graph Convolution
2D Predicted Keypoints of Probe and Hand
Ground-truth
2D Loss
2D Predictions
Adaptive Graph U-Net
3D Predicted Keypoints of Probe and Hand
Ground-truth
3D Predictions
3D Loss

Figure 5.3: HOPE-Net architecture for multi-modality input using RGB and depth images. The diagram is illustrated with a sample synthetic RGB and corresponding depth image from the extended dataset as input.

the advantages of the proposed method. Taking into account the above-mentioned potential advantages of multimodality, depth information was integrated into the HOPE-Net training and evaluation method, which was previously based exclusively on RGB, as described in the HUP-3D chapter 4.

Multi-modality configuration. The original HOPE-Net model, detailed in the paragraph on single-modality architecture and illustrated in Fig.5.2, estimates 3D hand-object poses from single RGB images and does not include additional depth information. Consequently, HOPE-Net is a single-modality neural network architecture that processes RGB frames using a single backbone[6]. To support a multi-modal approach and process additional depth frames, a second ResNet-18 backbone was added to the original HOPE-Net model to process depth frames independently. The resulting architecture is shown in Fig.5.3. The outputs of the two backbones—i.e., the feature representations from the RGB and depth input images—are concatenated and fed into the graph convolution layer. Additionally, each ResNet-18 backbone produces feature encodings that are passed to a separate 2D hand and probe keypoint predictor. The resulting 2D keypoint features, extracted independently from the RGB and depth modalities, were concatenated to create a fused

---

[6] A backbone can be considered the first stage of a deep learning data processing pipeline. Its task is to learn low- and high-level image features such as textures, edges, and more complex object shapes. In this sense, a backbone serves as a reusable foundation for subsequent downstream components such as pose estimation layers.

representation, which was then used as the input node features to the graph convolution network. This fused representation provides extra information because depth adds distance and shape cues and clean occlusion boundaries that RGB lacks. Concatenating the modality-specific 2D keypoint features yields node embeddings that combine appearance (RGB) and geometry (Depth), allowing the graph network to exploit whichever modality is reliable at each hand joint and probe bounding box corner. The remaining components of the model architecture—including the Adaptive Graph U-Net—are identical to those in the single-modality architecture presented in the paragraph on single-modality architecture. Since depth maps are purely synthetic—as detailed in Chapter 4—and rendered from the scene geometry (camera z-buffer), no physical sensor is used. Consequently, no sensor-specific noise is present, and RGB and depth are rendered at the same spatial resolution (256 × 256) before being resized to the network input. Injecting sensor-specific depth noise during Blender rendering is a possible enhancement, but is not considered in this work and left to future work.

5.3 Experiment: Pose Estimation Evaluation

Demonstration of the utility of the extended dataset HUP-3D-v2 follows the same experimental setup as for the HUP-3D case, Section 4.3 of Chapter 4. The major difference of the evaluation presented in this section is the additional use of a multi-modality setup that required a change in HOPE-Net's architecture, as presented in Section 5.2.1. The 63,360 image sets of the HUP-3D-v2 dataset comprise 22 realistic hand-object grasps, with each grasp contributing 2,880 image sets. To ensure the generalisability of the tested HOPE-Net model, the data were split into 14 grasps for training (40,320), four for validation, and four for testing (11,520). This follows a similar supervised learning setting as presented in Chapter 4, but with twice the number of image sets.

To ensure comparability with earlier evaluations, model training on HUP-3D-v2 used the same hyperparameters as the HUP-3D evaluation presented in Section 4.3 of Chapter 4. During testing, HOPE-Net was evaluated on the 11,520 test images using the same batch size as in model training (64), along with GPU acceleration and a pretrained model checkpoint obtained from prior training. The dataset's RGB and Depth images have a resolution of 256x256, consistent with the ResNet backbone's expected input scale. The actual evaluation of HOPE-Net on the aforementioned test images, in both single- and multi-modality modes using the two respective model architectures detailed in Section 5.2.1, is presented in Subsections 5.3.1 and 5.3.2.

5.3.1 Single-Modality Evaluation

Table 5.1 shows the performance of HOPE-Net on HUP-3D-v2, evaluated in both single- and multi-modality modes. For comparison, it also includes a baseline result: the original single-modality evaluation on HUP-3D, as presented in Chapter 4. In the single-modality case where only RGB frames were used—as in the HUP-3D evaluation described in Chapter 4—the addition of Clarius frames to the Voluson dataset slightly reduced the performance of the HOPE-Net model, resulting in an average MPJPE of 13.01 mm. This decrease is likely due to the increased complexity introduced by using more than one US probe: the newly added Clarius probe increases the overall diversity in terms of object shapes and sizes, as well as the interaction dynamics between hand and probe. Models must learn to appropriately incorporate this greater diversity to generalise effectively. Selected qualitative results for the single-modality evaluation on HUP-3D-v2 are illustrated in Figure 5.4. It can be observed that some frames show notable deviations between predictions and ground truth.

5.3.2 Multi-Modality Evaluation

The original HOPE-Net model, as presented in [51], and detailed in the paragraph on single-modality architecture, estimates 3D hand-object poses from single RGB images only and is therefore a single-modality model. The availability of depth images in the HUP-3D v2 dataset motivated an extension of the original architecture to incorporate a second modality, as described in the previous Subsection 5.2.1. Since RGB and depth data, denoted as RGB-D, provide mutually complementary

Dataset	**Model version**	**MPJPE (mm)**		
		Hand	**Object**	**Average**
HUP-3D (Voluson)*	**HOPE-Net on RGB**	**5.33**	**17.05**	**8.65**
HUP-3D v2 (+ Clarius)	HOPE-Net on RGB	9.61	21.93	13.01
HUP-3D v2 (+ Clarius)	HOPE-Net on RGB-D	8.45	23.91	12.71
HUP-3D v2 (+ Clarius)	ResNet-18 on RGB-D	52.93	61.85	55.39

Table 5.1: Pose prediction performance of HOPE-Net on different datasets and configurations, including a ResNet-18 baseline for comparison. Performance is evaluated using MPJPE, which measures the average Euclidian distance between predicted and ground-truth joint positions (in mm): *MPJPE – 3D Hand* reports the error for the 3D hand joints (21 coordinates), *MPJPE – 3D Object* reports the error for the 3D positions of the US probe's bounding box corners (8 coordinates), and *Average MPJPE* combines both to summarise the total error accross 29 prdicted 3D coordinates.

* **Note:** The results for *HUP-3D (Voluson)* in the first row were originally presented in Chapter 4 and are listed again for comparison.

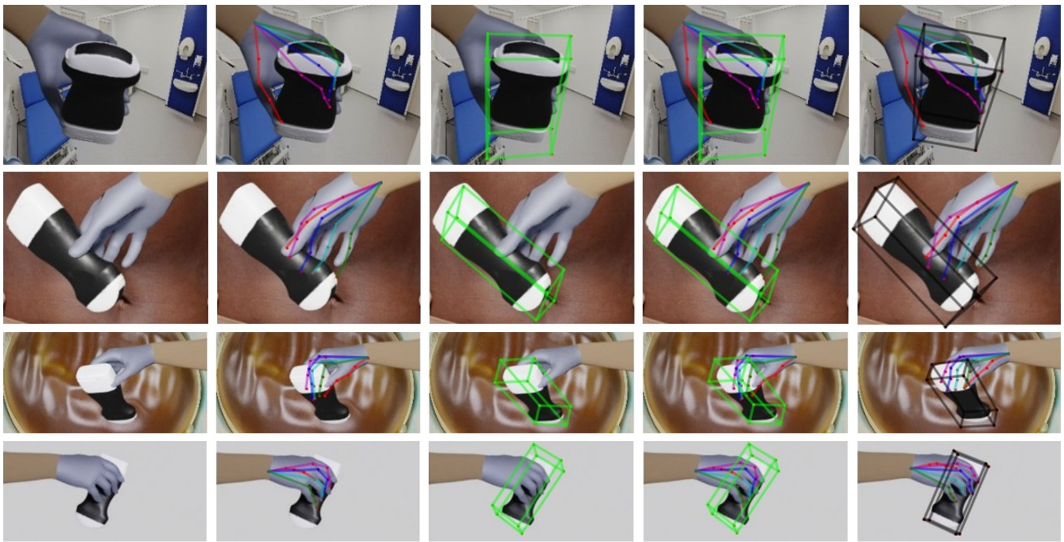


Figure 5.4: Qualitative results obtained from single-modality evaluation on HUP-3D-v2, shown using four test images from the dataset: image columns from left to right: RGB, predicted hand joints, predicted probe corners, predicted joints and corners, ground truth of joints and corners. Visual inspection shows notable deviations between predicted and ground truth joints and corners.

information, their combined use can improve the robustness of 3D hand pose estimation [291]. While RGB data provides color and texture information, depth data, through its spatial and geometric information, provides an additional complementary source of information that can help reduce ambiguities related to hand orientation and occlusion.

The pose prediction performance of the trained multi-modal HOPE-Net model on the test dataset is reported in the third row of Table 5.1. The hand and total average MPJPE errors are slightly lower than the ones of the HOPE-Net on RGB result shown in the second row of the table. However, the object error of the RGB-D result is slightly higher than the RGB only results. Overall, it can be concluded that the addition of a second modality in the presented setup did not lead to considerable improvements in terms of pose prediction accuracy. A comparative ResNet-18 baseline was also tested on RGB-D using the same dataset—as shown in the last row of Table 5.1—but showed significantly worse performance than HOPE-Net. The underlying reason for testing a simplified model—ResNet-18 in this case—was to see whether such simplification improves the pose prediction results compared to the more complex HOPE-Net model. Despite its simpler architecture, such a model must still learn a highly non-linear mapping due to the larger output space of 29 x 3 coordinates. In contrast, HOPE-Net offers the advantage of a two-stage approach: it first predicts the 2D keypoints and then refines them into 3D coordinates, effectively managing the complexity of this non-linear mapping. In this case, however, the simplified ResNet-18 baseline clearly showed inferior performance compared to HOPE-Net, indicating that it is not a suitable choice for end-to-end 3D hand and probe pose prediction on the HUP-3D v2 dataset.

Given the modest performance improvement of HOPE-Net in the RGB-D configuration compared to the RGB-only setup, it can be inferred that optimising the two ResNet-18 backbones—by minimising their respective loss functions to reduce prediction error—was less effective in accurately regressing hand and probe keypoints than initially anticipated. This performance decrease could potentially happen due to two reasons: Firstly, adding a second backbone increases the number of trainable model parameters such as layers, weights and biases. Consequently, this leads to a higher number of parameters involved in minimising the highly non-linear function of the keypoints distribution. Specifically, keypoint regression is about learning non-linear relationships between the input features extracted by the backbone and their corresponding keypoints. Secondly, simply concatenating RGB and depth features can cause feature entanglement: modality-specific cues—appearance from RGB and geometry/occlusion from depth—get mixed in a way that the network cannot process them independently, which reduces the informativeness of the fused latent representation and can degrade keypoint regression. In other words, the features become overly correlated, obscuring which modality contributes what. This makes it harder for the network to map modality-specific signals to the correct 3D keypoints.

These experimental observations suggest that simply increasing input modalities does not guarantee considerably improved model performance. While depth information is theoretically beneficial due to its geometric complementarity, effective integration into the existing HOPE-Net model architecture remains nontrivial. The challenges observed in this study highlight that naive concatenation or fusion of RGB and depth features may even degrade performance if the network fails to learn meaningful representations from both modalities. Future work may focus on more advanced modality fusion strategies or improved network architecture components that more effectively handle the complementary nature of RGB and depth data.

5.4 Discussion

The two-fold attempt to increase the dataset's diversity by adding Clarius US probe frames and to introduce a second modality presented several challenges that are briefly discussed in this section. These efforts aimed to improve model generalisability in hand-probe interaction data, aligning more closely with real-world variability that may occur in clinical training scenarios. However, such extensions also introduced new sources of error and complexity, which impacted the model's predictive performance and raised important design considerations for future iterations.

The first challenge, as described in section 5.2, is that the HUP-3D-Clarius dataset contains frames where both the right hand's thumb and index finger show slight interpenetration with the probe. This slightly reduces the realism of the synthetic images and may affect model performance. A potential solution to this interpenetration problem could be to retrain the grasp generating generative model that was presented in chapter 4, section 4.2.2 with the newly aquired HUP-3D-Clarius images to fine-tune its performance. Instead of retraining the model from scratch, it can be initialised from its last saved checkpoint to preserve its learned weights and biases, which reduces the overall training time and effort. A retrained model might be able to avoid interpenetration of the hand and the Clarius probe.

The second challenge lies in the fact that adding Clarius images to the dataset reduces HOPE-Net's ability to generalise, as described in Subsection 5.3.1, ultimately leading to a decrease in performance. To some extend, this could be related to the aforementioned interpenetration of hand and probe, which reduces the realism of the synthetic images. However, it is more likely that the model's inability to adapt to the greater variability introduced by the added Clarius images, and required for effective generalisation, outweighs the interpenetration issue.

Lastly, the third challenge emerged from the attempt to implement a multimodal approach using depth frames alongside RGB frames, as described in Subsection 5.3.2. Although this setup resulted in similar pose prediction performance compared to the RGB-only configuration, the simple concatenation of initial ResNet-18 feature encodings and 2D hand and probe keypoint predictions turned out to be a suboptimal architectural choice, indicating a need for further refinement.

Addressing these challenges is out of the scope of this thesis and hence part of potential future work. However, they offer valuable insights into current limitations of synthetic dataset design and model integration strategies. Addressing these issues in future research could lead to more robust and generalisable pose estimation models for medical applications.

5.5 Conclusion and Future Work

This chapter presented a continuation of the work presented in the previous Chapter 4, by focusing on two contributions: First, increasing the dataset's diversity by generating frames with a second US probe, the Clarius C3, resulting in a new standalone dataset called HUP-3D-Clarius and a dataset that combines HUP-3D and HUP-3D-Clarius, called HUP-3D-v2. Both newly introduced datasets are publicly available and aim to support further research in the area of markerless 3D hand and tool pose estimation in the to date underexplored area of pose estimation in obstetric US. The second contribution this chapter presented consisted of an attempt to evaluate HOPE-NET's performance on the extended dataset, to test potential benefits of increased dataset diversity and associated improved model generalisability when trained with these data. Evaluation was conducted in single-modality mode, like in Chapter 4, and also in multi-modality mode. Exploring potential benefits of multi-modality was a novel, yet unpublished contribution to this chapter. Supporting multi-modal input required a minor architectural modification to HOPE-Net. Although the results in dual-modality setup did not show a considerable improvement over the single-modality configuration, potential causes were discussed in Section 5.4, and further investigation is recommended. Future improvements to the multi-modality network architecture could involve replacing the simple concatenation of feature encodings with a weighted fusion method that assigns modality-specific importance, potentially enhanced by an attention[7]-based mechanism to dynamically prioritise characteristic features from each modality.

The inclusion of the Clarius wireless probe opens the door to AR-assisted medical education in low-resource settings, as it offers a more cost-effective alternative to conventional US systems like the Voluson E10. Another advantage of wireless probes is that the US image can be displayed on smaller, more portable devices such as smartphones, tablets, or compact monitors. These can be positioned closer to the patient's anatomy, allowing novice practitioners to monitor live US output without having to divert their gaze away from the scanning site. Such setups can support both actual sonographic procedures and prior MR-assisted training sessions.

Future work could explore a third modality the datasets provide: segmentation masks. This third image modality may help pose prediction models to learn hand

---

[7] Attention mechanisms are neural network components that learn to assign individual weights to different input features, allowing the model to focus on the most relevant feature representations dynamically.

regions, probe boundaries or probe contact areas more explicitly. In addition, feature localisation could be improved, which may be useful in occluded or cluttered scenes. Incorporating segmentation information could also facilitate better spatial disentanglement between hand and US probe during the early stages of the network architecture, potentially improving model robustness in complex interactions.

Another potential future work could be to increase the dataset's variety even further by including images with additional US probes. Such an effort, however, would depend on the availability of other conventional probes that are available in laboratory settings to be laser scanned and 3D reconstructed as models to be imported into Blender. A collaboration with hospitals could, for example, allow access to such other US scanners. Beyond adding real models, parametrically randomised probe geometries could be used that preserve the canonical form of commercial US probes while introducing local variations. This would increase image and grasp diversity without compromising clinical plausibility.

This chapter concludes the thesis's technical contributions. Chapter 6 presents the overall conclusions and directions for future work. The design, implementation, and evaluation of markerless hand-tool pose estimation lay a foundation for future research in medical training systems and contribute to the broader goal of developing robust, accessible, data-driven simulation tools. The presented methods may also translate to other domains (e.g., surgical procedures) where similar principles apply.

Before concluding in Chapter 6, Subsection 5.5.1 briefly describes a prototyping effort that explores an alternative way to extend the diversity of the synthetic dataset via simulated grasp movements.

5.5.1 Extending Data Diversity of HUP-3D Through Simulated Grasp Movements

The hand grasp generation method used to create plausible grasps for the HUP-3D and HUP-3D-Clarius datasets (as discussed in Chapters 4 and 5) had limitations in generating a greater diversity of plausible grasps. GrabNet [269], a pre-trained generative model that generates grasps based on hand-object contact areas learned from everyday objects, enables the creation of plausible grasps but lacks diversity in grasp contact areas and hand grasp configurations. Retraining the model with images of hands grasping US probes could improve the plausibility of generated grasps, but might still limit the diversity of plausible grasps.

Another method of finding plausible grasps that has recently been explored in the research community is the use of a physics simulation engine for robotics and AI named RaiSim[8] [294] to generate plausible grasp motions [295, 296, 297]. RaiSim, a rigid-body simulator, is designed for accuracy and speed in simulating robotic systems and has been used for RL-based robot control simulations [298, 299, 300]. In RL, unlike traditional DL, no ground truth is required for model learning. Instead, an agent—such as a robot—explores a simulated world and autonomously learns the connections between inputs and outputs through states, actions, and rewards. Each step closer to the optimal state results in a reward, while steps further away lead to a penalty. The application of suitable reward policies is crucial in this context. However, one downside of this approach is that RL often requires a significantly high number of trial-and-error runs before achieving an acceptable training state. On the other hand, when applied in suitable simulation environments, RL algorithms are able to offer policies that generalise well to unseen situations [301]. While such RL-based grasp synthesis does not guarantee human-like biomechanics, it yields full approach trajectories (not just final hand poses), enabling pre-contact samples and greater diversity than static, DL-based grasp generators.

In [296], a RL-based hand-object grasp motion generation concept utilising the RaiSim simulation engine is presented that builds on previously developed similar concepts [302, 295], but offers improved scalability to unseen grasp objects. More specifically, [296] proposes a RL policy learning framework called *GraspXL* capable to generate a wide variety of grasp motions, including grasp motion objectives and hand morphologies. GraspXL allows users to define the hand's heading direction towards the graspable object and the wrist rotation. Based on a predefined user-specific graspable area of the object, the target midpoint and target heading

---

[8] `https://raisim.com/index.html`. Accessed 27 January 2025

direction originating from the target midpoint can be defined, allowing flexible control and greater variability in terms of the generation of plausible grasps. Compared to the generative DL model-based grasp rendering method presented in chapter 4, GraspXL offers two key advantages: (i) greater flexibility in generating a wider variety of plausible hand grasps, and (ii) the ability to simulate complete grasp motions, including the hand's approach phase prior to contact with the object. The second advantage could enhance the HUP-3D grasp generation method by enabling the extraction of hand-object poses at earlier stages of the motion, specifically where the hand has not yet made contact with the object. Adding rendered synthetic images of hands, depicting interval-based fragments of entire grasp motion sequences, to the HUP-3D dataset could potentially enhance the generalisability of pose estimation networks like HOPE-NET when trained on such datasets. This could also introduce a level of novelty to the research community, as pre-grasp images of hands in close proximity to objects do not appear to be included in existing benchmark datasets such as [256, 251, 280, 240, 241, 257].

In an initial feasibility attempt to generate grasp motions using RaiSim, the hand grasp simulation codebase *D-grasp*[9] [295], a predecessor of *GraspXL*, was installed on a computer equipped with a 13th Gen Intel® Core™ i9-13900H processor (2.60 GHz), 64 GB of RAM, and running Windows 11 Home (64-bit). However, attempts to run the provided hand grasp motion demos—featuring Unity[10] game engine visualisations connected to a Python-based grasp simulation backend via TCP/IP sockets—were unsuccessful. The visualisation did not display any hand grasp motions, as the application became unresponsive while attempting to send RaiSim-based motion updates to the Unity engine via TCP/IP. It is hypothesised that this issue may stem from compatibility limitations in the older RaiSim version used by *D-grasp*, particularly when deployed on Windows systems using socket-based communication.

Following the unsuccessful attempt with *D-grasp*, the successor framework *GraspXL* was explored. The author of [296], Zhang Hui, was contacted and pro-

---

[9] `https://github.com/christsa/dgrasp`. Accessed 27 Jan. 2025

[10] `https://unity.com/`. Accessed 27 Jan. 2025

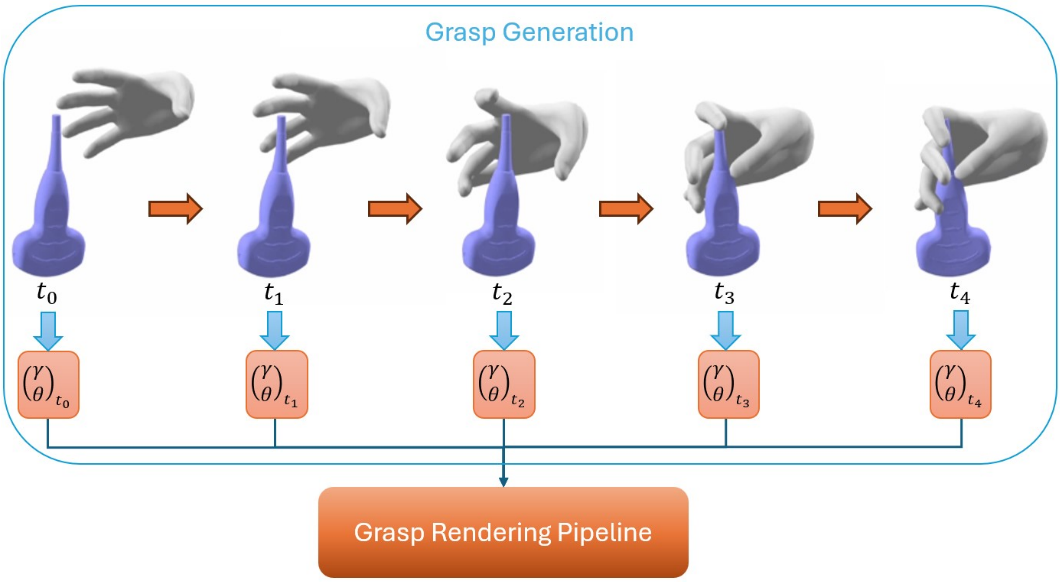

Grasp Generation
t_0
t_1
t_2
t_3
t_4
$\binom{\gamma}{\theta}_{t_0}$
$\binom{\gamma}{\theta}_{t_1}$
$\binom{\gamma}{\theta}_{t_2}$
$\binom{\gamma}{\theta}_{t_3}$
$\binom{\gamma}{\theta}_{t_4}$
Grasp Rendering Pipeline

Figure 5.5: Schematic illustration of a potential future grasp rendering pipeline that leverages the capabilities of grasp simulation engines: interval-based hand position samples $(\gamma = hand\ \ position, \theta = hand\ \ pose)_{t_i}$ of an entire grasp motion sequences, recorded at five time stamps $t_i \in [1, \ldots, 4]$

vided with a model of the Voluson US probe. Using the *GraspXL* simulation engine, sample grasp motion simulations were generated and delivered as `.npy`[11] data files, along with a video recorded from the RaiSim-based simulation. Fig. 5.5 shows five representative grasp states of the Voluson probe at timestamps $t_i \in [1, \ldots, 4]$ extracted from the video.

In addition, Fig. 5.5 illustrates the envisioned grasp generation and grasp rendering pipeline, leveraging the capabilities of the *GraspXL* simulation engine: Various hand poses $(\gamma = hand\ \ position, \theta = hand\ \ pose)_{t_i}$ including hand position γ and hand joint configuration θ are extracted at sampled time stamps $t_i \in [1, \ldots, 4]$ throughout the grasping motion simulation. The hand poses poses $(\gamma, \theta)_{t_i}$ can then be fed into the Blender 3D software-based grasp rendering pipeline I proposed in chapter 4. Following an initial feasibility test in which data from a single hand pose $(\gamma, \theta)_{t_i}$ was extracted and mapped to the hand pose parameters required by the HUP-3D grasp renderer, the resulting rendered Blender image displayed an incorrect hand pose, as shown in Fig. 5.6.

---

[11] An .npy file is a file format used by *NumPy*, a Python programming language library commonly used for numerical computations in Python.

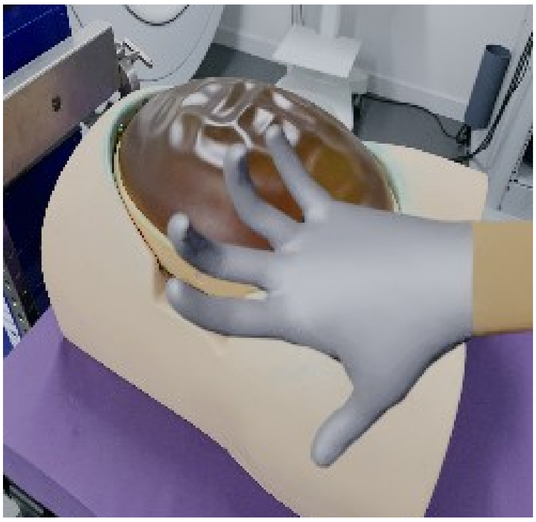


Figure 5.6: Synthetic RGB example image of a hand with mismapped hand pose: Failed attempt to map an example hand pose generated by a *GraspXL* hand grasp simulation sequence by Hui Zhang [296] into the Blender grasp rendering pipeline presented in chapter 4. Correct mapping of the MANO hand model parameters from SMPL-X [303] to the SMPL-H [276] body model is left for future work.

The reason for this incorrect 1:1 mapping of the hand pose information lies in the fact that *GraspXL* uses the SMPL-X [303] representation of the human body, while the HUP 3D grasp renderer uses the older SMPL-H [276] representation of the human body. The representations of the 48 MANO hand model parameters are different in SMPL-H and SMPL-X. SMPL-X's GitHub repository contains model parameter transfer code[12] that allows users to transfer a SMPL-X model to a SMPL-H model. Converting SMPL-H to SMPL-X using this model transfer code could potentially resolve the aforementioned issue of incorrect MANO hand pose mapping; however, this is left for future work.

In summary, using a grasping motion simulation engine such as *RaiSim* together with a RL policy learning framework as presented in *GraspXL* [296] could be a viable option to extend the current HUP-3D dataset concept. On the one hand, *GraspXL* in particular offers more flexible ways to define hand grasp positions via hand motion direction, hand orientation, and object contact areas. This could lead to the generation of a larger variety of plausible grasps. On the other hand, *GraspXL*

---

[12] `https://github.com/vchoutas/smplx/tree/main/transfer_model`. Accessed 27 Jan. 2025

could be used to generate additional images showing hands approaching the US probes and would potentially increase the diversity of a dataset. And datasets with higher image diversity will certainly improve the generalisability of 3D hand pose estimation networks such as HOPE-NET when trained on such datasets.

Chapter 6

Conclusions and Future Work

6.1 Summary of Contributions and Limitations

As this thesis concludes, this section summarises its main contributions and limitations, as detailed in the following subsections. The main contributions—each of which has been published in a peer-reviewed article—include: a literature review of OST-HMD-assisted surgery; the development of an innovative MR application for training in obstetric sonography; and a novel concept for markerless joint 3D hand and US probe pose estimation. These contributions span the domains of OST-HMD-based human-computer-interaction, medical training, computer vision, and synthetic data generation for DL-based 3D pose estimation, and collectively aim to improve usability of MR systems for clinical and educational settings.

Each contribution addressed specific research challenges. The literature review chapter focused on categorising technological and human factor trends to identify current possibilities and limitations, and propose considerations for future system design. The CAL-Tutor MR prototype introduced a novel training platform with promising results in early usability testing. The markerless 3D pose estimation pipeline explored scalable synthetic image dataset generation; and demonstrated strong baseline pose estimation performance using RGB input.

Chapter 5 contains unpublished work. Subsection 6.1.4 summarises the continuation of the work first presented in Chapter 4 and later extended in Chapter 5. This extension focused on broadening the dataset scope and investigating the impact of multi-modality input for 3D pose estimation. While preliminary results offer valuable insights, they also underscore the complexity of achieving robust generalisability in DL models trained on synthetic data. At the time of writing, this work is intended for future publication, contributing to the growing research interest in markerless hand-tool pose estimation in medical contexts.

The contributions outlines in the following subsections represent a clear progression from in-depth critical review and MR system prototyping to the development of scalable data generation pipelines for training DL-based pose estimation models. This trajectory reflects a shift from understanding the current literature landscape and identifying design gaps, to creating practical systems and resources that can directly innovation in clinical training environments. Their respective limitations are also discussed to highlight technical challenges, usability constraints, and areas where further investigation is required. By doing so, this thesis aims not only to contribute new tools and methods but also offer guidance for future research efforts in this evolving and still underexplored field.

6.1.1 Systematic Review on AR in Surgical Applications

The systematic literature review of OST-HMD-assisted surgical applications, presented in Chapter 2, represents the first contribution of this thesis. It involved categorising 91 studies published between 2013 and 2020 by application domain, technologies used, research settings, and human factors. The review revealed key trends in the field, the significant upward trend in published articles, as well as technical and human factor limitations that hinder widespread clinical adoption. The chapter is based on a corresponding publication [53], which provides one of the most in-depth analyses of OST-HMD applications in surgical contexts published to date. The identification and categorisation of key human factor limitations at the individual study level represents a novel contribution of this review and supports the claim that human factors emerge as a significant determinant of OST-HMD utility. This review highlights that future success in OST-HMD-assisted surgery depends not only on addressing technical and perceptual challenges in controlled environments, but also on integrating human factors into system design to solve specific clinical problems. This combined, problem-driven approach represents a key strategic recommendation of this work.

One of the main findings of the literature review was that the clinical adoption of safety-critical OST-HMD-based surgical applications—such as surgical guidance—requires consistently high registration accuracy of 3D virtual objects and perceptual accuracy, which is difficult to achieve in practice. This insight contributed to a shift in focus of this thesis toward clinical areas where precise 3D virtual overlay is less critical, such as medical education. As a result, the CAL-Tutor MR system was developed, as discussed in Subsection 6.1.2 below.

6.1.2 Mixed Reality for Medical Training

A focus shift from surgical AR to AR-assisted medical training motivated the work presented in Chapter 3, and led to the second contribution of this thesis: the development of the CAL-Tutor MR system for training in obstetric sonography. CAL-Tutor addressed a gap in existing MR training applications by focusing on prenatal US and was designed to shorten the learning curve for US trainees and improve spatial understanding of US probe placement and orientation. In particular, CAL-Tutor was designed to support the fine motor coordination and spatial reasoning skills essential for accurate fetal biometry in obstetric US training.

The system demonstrated the feasibility of combining a real US scanner with a phantom mother's belly and fetus, using an OST-HMD to deliver in-situ 3D virtual guidance. An initial user study involving engineers without prior knowledge of MR and obstetric sonography found that the MR guidance improved several aspects of system interaction, such as efficiency, clarity and stimulation. A key characteristic of the CAL-Tutor platform is its ability to record user motion data by leveraging the capabilities of the HoloLens 2. It is assumed that this data has the potential to provide meaningful insights into behavioral motion patterns distinguishing novice and expert users. An initial evaluation of visual attention profiles of the study participants using recoded eye gaze data reveals interesting insights, but a higher population size would be required to obtain meaningful statistical insights. The MR source code of CAL-Tutor is publicly available to support reproducibility and encourage further work in the area of MR-assisted medical education; especially in the underrepresented field of training in obstetric sonography.

A key limitation of the CAL-Tutor system was the limited accuracy of virtual overlay alignment and US probe tracking, due to its ArUco marker-based tracking approach. Similar challenges have been reported in recent work on MR training for obstetric US, where marker-based probe tracking was also found to be a limiting factor [229]. This limitation motivated the third contribution of this thesis, which focuses on markerless 3D hand and probe pose estimation; summarised in the following Subsection 6.1.3.

6.1.3 Synthetic Dataset and Single-Modality Evaluation for 3D Hand–Probe Pose Estimation

The limitations of marker-based tracking of the US probe tracking within the CAL-Tutor system, discussed in previous Secion 6.1.2, shifted the focus of this thesis towards a markerless approch to obtain the probe's 3D pose, presented in Chapter 4. In addition, 3D hand pose estimation was considered alongside probe pose estimation due to the relevance of both in markerless tracking for MR applications. Consequently, the third contribution of this thesis focused on a novel approach for markerless joint 3D hand and US probe pose estimation. The underlying idea was to focus on the generation of training images for deep learning-based, end-to-end pose estimation models.

Contributions focused on a scalable synthetic multi-modal image generation pipeline that combined a generative model for grasp generation and a 3D computer graphics software for grasp rendering. A novel sphere-based camera viewpoint concept rendered grasp images from egocentric and non-egocentric viewpoints and thereby enhanced frame generalisability. These efforts resulted in a synthetic, multi-modal image dataset named HUP-3D, which includes RGB-D images and segmentation maps of hand grasps performed with surgical gloves, using an US probe commonly used in obstetric sonography, the Voluson C1-5-D US probe. Single-modality evaluation (RGB only) of HUP-3D on a trained state-of-the-art model, HOPE-Net [51], showed lowest hand and object 3D pose estimation errors.

The HUP-3D dataset, along with its synthetic data generation pipeline and evaluation code, is publicly available via a dedicated project website to support reproducibility and encourage further research in the area of clinical synthetic data generation and markerless 3D hand and tool pose estimation. While the initial synthetic dataset focused on a single US probe and single-modality input (RGB), these constraints limited the generalisability and scope of the evaluation. To overcome these limitations, the dataset was expanded to include a second probe type and multi-modal data, and a comprehensive evaluation was conducted. These advancements are summarised in the following Subsection 6.1.4.

6.1.4 Dataset Extension and Multi-Modal Evaluation for 3D Hand–Probe Pose Estimation

The extension of the HUP-3D dataset, presented in Chapter 5, marked the last contribution of this thesis. The dataset got extended with synthetic images of hands holding a second, wireless US probe, the Clarius C3, representative of modern, portable US technology. The same grasp generation and image rendering method used for creating HUP-3D was applied. Using the Clarius probe, however, revealed a limitation of the generative model-based grasp generation method. The larger size of the Clarius requires the hand to open wider, resulting in minor interpenetration between the probe and the fingers. However, since the extent of this interpenetration remains minimal, retraining the generative model to accommodate the Clarius's larger dimensions and generate more precise grasps with less interpenetration was considered disproportionate in terms of time and effort required. The extended dataset, referred to as HUP-3D-v2, comprises an equal number of Voluson and Clarius grasp images and was evaluated using HOPE-Net, as detailed in Chapter 4. In addition to a single-modality evaluation using RGB images only, a multi-modal evaluation incorporating both RGB and depth frames was conducted. While the dual-modality setup yielded performance results comparable to the single-modality configuration, the overall hand and probe position errors in HUP-3D-v2 were higher than those in HUP-3D—likely due to the model's limited ability to generalise to the increased variability introduced by the inclusion of a second probe. Optimising the proposed HOPE-Net dual-modality network architecture is left for future work.

6.2 Future Work

The work presented in this thesis opens the door to several potential areas for future research. While Chapters 3 to 5 each offer contribution-specific suggestions, this section outlines broader directions beyond the scope of the individual chapters. These directions reflect both the practical implementation challenges encountered during system development and the conceptual opportunities identified through evaluation and experimentation. Together, they aim to inspire future efforts to design intelligent, adaptable, and scalable solutions for medical training using MR and ML.

CAL-Tutor offers multiple opportunities for further development, particularly in response to practical limitations and emerging needs in global healthcare education. A potential high-impact application context is the integration of portable US scanners to support more cost-effective US training in low-resource or remote settings, helping bridge the gap in sonography education worldwide. Additional enhancements could include the incorporation of DL-based MR guidance to assist users in acquiring standard US planes, as described in Subsection 6.2.1.

The HUP-3D data generation pipeline also presents several future research directions that aim to improve the realism and effectiveness of synthetic training data. More advanced grasp synthesis techniques, particularly those involving high-fidelity generative models capable of incorporating more effective hand grasp positions and orientations, as well as tool-related contact regions, could produce hand-tool interactions that better reflect the dexterity and variability of real clinical tasks. Similarly, enhancing background realism through domain-specific textures of simulated clinical environments could improve model generalisation, especially for ML-based pose prediction tasks. Building on Chapter 5, future work could add a third modality—segmentation masks—to complement RGB-D. This semantic layer would help models localise hands, probe boundaries, and contact regions, improving model performance under occlusion and visual clutter. In turn, it could yield more robust 3D pose estimation for autonomous or semi-autonomous training and assessment tools.

Complementary future work suggestions not covered in the contribution chapters are outlined in Subsection 6.2.1.

6.2.1 Further Development of the CAL-Tutor MR Application

The CAL-Tutor MR application, detailed in Chapter 3, introduced an initial prototype aimed at enhancing medical education in fetal US, while leaving several opportunities for further development. Foremost among these is the integration of the proposed markerless hand and probe tracking, as presented in Chapter 4. The central idea is to employ a trained, state-of-the-art 3D pose estimation model, such as HOPE-Net, to enable continuous, real-time pose estimation during the use of the CAL-Tutor application. Given the limited processing capabilities of the HoloLens 2, real-time model inference may need to be offloaded to an external computer. In addition to integrating markerless pose estimation, recording and evaluating user motion data should also be considered. The current version of CAL-Tutor captures limited hand pose data—specifically palm and wrist positions—using the built-in HoloLens 2 hand tracking. However, the proposed markerless approach could enable more detailed finger tracking, even under mutual occlusion of the hand and probe. The fact that the pose estimation model is trained using both egocentric and non-egocentric images may offer significant advantages in such occluded scenarios.

The recorded user motion data—including probe and hand poses, eye gaze, and an external video feed—could then be analysed to gain valuable insights into behavioral differences between novice and expert users. Further experiments would be required to explore this in detail. Moreover, the collected data could be leveraged to train other ML models capable of inferring user intent and delivering real-time, context-aware 3D virtual assistance. Currently, CAL-Tutor displays persistent AR guidance until the user has successfully navigated the probe to the respective target US planes. However, this permanent guidance may not be ideal for all novice sonographers, as it could make navigation appear too straightforward. Instead, a context-aware guidance system—capable of interpreting the user's motion patterns and providing assistance only when needed—may offer a more effective learning experience and improved educational outcomes.

A fundamental concept that was outside the scope of this thesis, but should be explored in future work, is the integration of the HUP-3D markerless 3D hand and tool pose estimation method into the CAL-Tutor application. Such integration would enable real-time markerless tracking of the US probe, eliminating the need for an ArUco marker cube attached to the US probe. An initial feasibility assessment of this integration was envisioned during the thesis but was ultimately abandoned due to technical limitations associated with the HoloLens 2's internal camera system. One major issue was the inability to access RGB and AHAT (high-frequency near-depth) data streams simultaneously. At some point, updates to the HoloLens 2's Windows 10–based operating system discontinued support for simultaneous RGB and AHAT depth acquisition, making RGB-D data collection possible only by downgrading the OS—which was considered an impractical solution. Another challenge involved the quality of the internal RGB camera. Visual inspection of captured frames revealed that the image quality was insufficient for reliable use in markerless 3D hand and tool pose estimation. A practical mitigation is to use an external RGB-D sensor (e.g., Azure Kinect DK (Microsoft, Redmond, USA) or a comparable device) rigidly calibrated to the HoloLens 2. This would require (i) rigid mounting and extrinsic calibration to the HoloLens coordinate frame, (ii) low-latency power/data and streaming to an external computer for inference, and (iii) attention to ergonomics and cable management. Given the discontinuation of Azure Kinect DK [304], suitable alternatives—e.g., Orbbec Femto (Orbbec, Troy, MI, USA))—should be evaluated in future work.

Despite these limitations, integrating the HUP-3D method into CAL-Tutor remains a promising direction for future work. This could potentially be achieved using external cameras that offer higher-quality RGB-D capture and are not constrained by the HoloLens 2's hardware limitations. Such external setups would also allow for more flexible placement and potentially wider field of view, improving the robustness of hand-tool tracking in dynamic training scenarios. Future implementations could benefit from compact, portable computing units—such as embedded GPUs or wearable AI accelerators [305]—and next-generation commercially available OST-HMDs, enabling seamless integration of high-quality RGB-D tracking into MR workflows without compromising mobility or performance.

6.3 Reflections and Broader Impact

This section reflects on the contributions, challenges and future research directions of the thesis, while also offering personal impressions from the author. The research presented spans three main areas, each with its own opportunities and challenges, yet all driven by a shared goal of generating meaningful impact. Together, these efforts contribute to advancing technology for medical MR, with a particular focus on training and education. Some of the methods and tools explored may also prove beneficial in surgical MR applications, particularly in scenarios where precise hand motions and mental mapping of 2D imaging information to 3D patient anatomy are crucial, such as surgical guidance. MR could support not only visual overlay but also guide a surgeon's hand movements—shifting the focus from traditional AR-assisted navigation to dynamic motion guidance. In such contexts, accurate tracking of hand-tool interactions is essential to ensure reliable and intuitive MR system performance.

To pursue the broader vision of advancing medical AR, the thesis was structured around a series of exploratory and practical contributions, beginning with a critical review of surgical AR technologies. First, this thesis aimed at exploring the state-of-the-art of surgical AR, with the initial intention in mind to create a MR solutions that improve surgical workflows within the scope of this thesis. However, the literature review (Chapter 2) revealed that demonstration of clinical utility is rare. One study even found that perfect registration can lead to unwanted side effects during guidance tasks, and that overlaid AR was found to cause inattention blindness [69]. One option for the further development of this thesis could have been to explore such a side-by-side but not directly overlaid 3D virtual visualisation for surgical guidance tasks. The emphasis on human factor limitations is certainly a novel approach on conducting a systematic literature review and hopefully inspires future research to categorize and address such human factors in more detail when designing surgical AR applications. The literature review turned out to be more in-depth and more complex and time consuming than initially planned. However, the fact that the respective publication [53] is well-cited indicates that this literature review was a relevant body of work for the research community.

Secondly, by focusing on the use of AR for medical education rather than surgery — with the CAL-Tutor application (Chapter 3) as the central MR system — this work explored a comparatively underrepresented area within the research community. As of this writing, articles exploring MR-based training systems specifically for obstetric US are rare, with the system proposed in [229] representing the closest approach to CAL-Tutor. The MR aspect of CAL-Tutor was not without limitations. The HoloLens 2 requires re-calibration for each user, and unwanted shifts of 3D virtual overlays can occur during use. Additionally, the registration of the 3D virtual object to the mother's abdomen—including the fetus—had to be performed manually, which is both time-consuming and error-prone. Moreover, the limited battery life of the HoloLens 2 further impacts usability. Future iterations of the application should aim to incorporate an automated registration process to improve both accuracy and user experience. The initial experimental results of the CAL-Tutor application suggest that MR holds promise for improving spatial understanding and reducing the learning curve in obstetric US training. However, further research with larger participant groups and extended evaluation periods is needed to confirm these findings. A combination of MR-based training and ML methods has the potential to significantly increase interest in this field, particularly by enabling intelligent 3D virtual feedback and user-specific learning assistance. Overall, MR-assisted training in obstetric US could offer a more cost-effective alternative to conventional simulators, potentially enabling broader access to critical medical education in regions where traditional training systems are financially out of reach.

Thirdly, the second focus shift of this thesis towards markerless 3D hand and probe pose estimation marks another niche area in the broader landscape of MR-assisted applications in obstetric sonography. While markerless hand and tool pose estimation methods are an active research area, most methods either focus on grasping everyday objects [256, 269, 251] or surgical tools [240, 257, 241]. US probes, however, come with specific requirements in terms of plausible grasp poses, which introduce additional challenges to automated grasp generation methods. In this context, the focus on markerless pose estimation for US probes presents a promising and underexplored research direction that should be explored in greater depth. The concept presented in Chapters 4 and 5, including the highly scalable synthetic image rendering method based on sphere-based camera viewpoint sampling, holds strong potential for advancing markerless 3D hand and tool pose estimation. By enabling the efficient generation of diverse and realistic ML training data, this approach can support broader applications beyond obstetric US, including surgical contexts. The rendering time and visual quality of the synthetic images, as well as the overall software-based rendering workflow, can certainly be improved. The current approach relies on a standard version of the 3D graphics software Blender [278]. However, a more advanced variant, BlenderProc2 [306], has been developed specifically for photorealistic rendering. It includes built-in functionality for generating realistic synthetic images and has already been successfully applied in surgical contexts [257]. Future iterations of the HUP-3D concept could incorporate tools like BlenderProc2 for improved image rendering. Overall, special emphasis was placed on maintaining low methodological complexity to enhance reproducibility and encourage further research in this area.

Fourthly, the use of RL-based grasp motion simulations, as described in Section 5.5.1, may open up new possibilities for generating realistic and diverse hand grasp poses, including pre-grasp configurations that enrich and diversify the dataset. Such enhanced image datasets could lead to improved ML model training and, ultimately, to more effective 3D hand and tool pose estimation. In particular, future work could explore integrating these datasets into real-time applications, bridging the gap between simulation and clinical environments. This thesis aims to encourage further research in this area.

6.4 Final Conclusions

This thesis, situated at the intersection of AR, 3D pose estimation within computer vision, and ML, presented three main contributions: (i) a comprehensive literature review of the state-of-the-art in surgical AR, with a particular focus on OST-HMDs, which identifies key gaps in the existing research and proposes directions for future exploration, particularly in addressing human factor limitations; (ii) the development of an AR application designed for training in obstetric sonography, leveraging the potential of MR in a niche field and encouraging further research in this area; and (iii) the introduction of a novel and scalable method for markerless 3D hand and tool pose estimation, with a specific emphasis on generating scalable and diverse synthetic image datasets for the specialised domain of obstetric US. Despite specific challenges such as hardware and software limitations of the HoloLens 2, and the generation of plausible grasp poses for US probes with DL models trained on everyday objects, the results demonstrate the potential of integrating AR, 3D computer vision, and ML to enhance medical education and, potentially, future surgical workflows. Looking ahead, further research into MR- and ML-assisted medical training methods may facilitate clinical adoption and support cost-efficient education, particularly in regions with limited infrastructure and financial resources. Additionally, the techniques developed in this thesis may serve as a foundation for advancing AR-guided surgical procedures, where the challenges and benefits of these technologies strongly align with those encountered in medical training.

This concludes the contributions and findings presented in this thesis, which span the intersection of AR-assisted surgery and medical education, the generation of synthetic hand–probe grasp images, and the use of DL-based generative models. It is the author’s hope that this work not only addresses existing gaps but also lays a foundation for more scalable, accessible, and intelligent solutions in AR-assisted medical training and guidance. As the technology continues to evolve, such interdisciplinary efforts may contribute to shaping the next generation of clinical education and surgical support systems.

Appendix A

Additional Details on the Literature Review of Chapter 2

This Appendix aims to provide some additional information in the context of the systematic literature review of OST-HMD-assisted surgery, presented in Chapter 2. The three tables shown below allow the reader to delve deeper into some information extracted from the 91 selected articles that were used for the in-depth analysis of the literature's state-of-the-art. Table A2 lists the 91 studies by the type of OST-HMD used, surgical context, and surgical procedure. The surgical context refers to the interventional setting, such as surgical guidance or preoperative planning, while the surgical procedure describes the actual intervention—such as needle biopsy or intraoperative bone localisation—that potentially benefits from OST-HMD assistance. Moving toward a crucial part of the systematic literature review—the identified human factors—Table A3 lists all addressed and persistent human factors per study. Addressed factors are limitations the OST-HMD aimed to overcome; persistent ones remained despite its use. Lastly, Table A4 provides an overview of the types of AR visualisation used, the conducted experiments, and the reported accuracy results. The table aims to establish a loose connection between the visual perception of 3D virtual content, the experimental evaluation of utility, and the types of measured accuracy, including reported accuracy values. The table thereby provides the reader with an overview of how utility was measured in relation to specific types of 3D virtual assistance.

Table A1: Acronyms used in Table A2 (surgical context)

SG	Surgical Guidance	PS	Preoperative surgical planning	SA	Intraoperative Surgical Anatomy Assessment
ST	Surgical training	TELC	Teleconsultation during surgery	REV	Intraoperative review of preoperative 2D imaging and/or patient records
TELM	Telementoring	DOC	Intraoperative Documentation	PM	Patient Monitoring
TP	Surgical Tool Placement	IO	Image overlay for navigation	SI	Screw Insertion
NI	Needle Insertion	CI	Catheter Insertion	KWI	K-Wire Insertion
EG	MIS Endoscopy Guidance	SP	Stent-graft Placement	DTG	Drill Trajectory Guidance
PN	Imaging Probe Navigation	SNN	Surgical Saw Navigation	CA	C-arm Positioning Guidance
RP	Robot Placement	DG	Dissection Guidance	AI	Anatomy Identification

Table A2: Studies listed by OST-HMD, surgical context and surgical procedure

Study	OST-HMD	Surgical context	Surgical procedure
[87]	Google Glass	TELC	Reconstructive limb salvage procedures
[88]	Google Glass	TELM	shoulder arthroplasty
[89]	nVisor ST60	SG (SI)	Percutaneous implantation of sacroiliac joint screw
[71]	Custom Device	SG (DTG)	Dental implant surgery
[163]	Google Glass	REV, ST, DOC, TELC	Different urological surgical procedures
[164]	Google Glass	ST, TELM	Inflatable penile prosthesis placement
[145]	Google Glass	PM	Bronchoscopy
[90]	nVisor ST60	SG (SI)	Percutaneous implantation of sacroiliac joint screw
[91]	Brother AirScouter WD-100G	SG (TP)	General intra-operative guidance (no concrete application, only measurement of attentiveness to the surgical field)
[131]	Moverio BT-200	SG (CI)	Central venous catheterisation under US guidance
[122]	Google Glass	SG (SI)	spine instrumentation (pedicle screw placement)
[93]	Google Glass	REV, TELC	Orthopaedic procedures
[149]	HoloLens	PS, ST	Preoperative diagnosis & planning of coronary heart disease
[132]	HoloLens	SG (CI)	Interventional endovascular stenting of aortic aneurysm
[66]	HoloLens	SAA, TELC	Visceral-surgical interventions
[157]	Moverio BT-200	SAA	Improvement of the body surface contour in plastic surgery.
[92]	PicoLinker glasses	SG (KWI)	Fluoroscopy controlled K-wire insertion into femur
[150]	Custom Device	PS	Preoperative diagnosis of coronary heart disease
[96]	HoloLens	SG (NI)	Percutaneous vertebroplasty, kyphoplasty and discectomy procedures
[97]	HoloLens	SG (KWI)	Percutaneous orthopaedic surgical procedures
[146]	HoloLens	SG (DTG)	Access cavity Preparation in Endodontic treatment
[98]	HoloLens	ST	Hip arthroplasty
[166]	HoloLens	SG (RP, TP, EG)	Increase the First Assistant's task performance during robot-assisted laparoscopic surgeries
[95]	HoloLens	SG (TP)	Intra-operative bone localisation
[123]	HoloLens	PS	Identification of a hemodynamic scenario that predicts an aneurysm rupture
[113]	HoloLens	SG (NI)	Needle biopsy
[124]	HoloLens	SG (IO)	Neurosurgical applications
[158]	HoloLens	SG (DG)	Vascular pedunculated flaps of the lower extremities (reconstruction surgery)
[94]	HoloLens	SG (CA)	percutaneous orthopaedic procedures
[115]	HoloLens	ST	example: transesophageal echocardiography examination
[114]	HoloLens	SG (IO)	N/A
[140]	Google Glass	DOC	Surgical time-out checklist execution
[138]	HoloLens	SG (TP)	N/A
[99]	HoloLens	SG (SI)	pedicle screw placement
[151]	HoloLens	PS	Repair for complex congenital heart disease
[100]	HoloLens	SG (IO)	Orthopaedic surgery (no specific procedure)

(continued on next page)

Table A2 (continued)

Study	OST-HMD	Surgical context	Surgical procedure
[103]	HoloLens	SG (KWI)	C-arm fluoroscopy guided k-wire placement
[101]	HoloLens	SG (AI)	Identification of spinal anatomy underneath the skin
[62]	HoloLens	SG (IO)	General image-guided surgical navigation (no specific application)
[102]	HoloLens	SG (SI)	Placement of pedicle screws in spinal fusion surgery
[152]	HoloLens	SG (CI)	transcatheter procedures for structural heart disease
[116]	HoloLens	TELM	Abdominal incision
[135]	HoloLens	TELM	Leg fasciotomy
[117]	HoloLens	SG (TP)	liver tumor puncture
[161]	HoloLens	SG (IO)	Head and neck tumor resections
[133]	HoloLens	SG (NI)	Seed implantation thoracoabdominal tumor brachytherapy
[139]	HoloLens	SG (IO)	General SG (no specific surgical application)
[126]	HoloLens	SG (TP)	Craniotomy
[127]	HoloLens	SG (NI)	Needle-based spinal interventions
[168]	HoloLens	PS	Nephron-sparing surgery
[104]	HoloLens	SG (TP)	Percutaneous orthopaedic treatments
[134]	Hololens	SG (SP)	Endovascular aortic repair
[147]	Magic Leap One	SG (PN)	Tooth decay management
[160]	Moverio BT-200	SG (SSN)	Mandibular resection
[172]	Vuzix M300	PM	None
[167]	HoloLens	RP	Set up of robotic arms by surgical staff (especially minimally invasive gastrectomy (abdominal surgery))
[118]	HoloLens	PS	Liver resection
[125]	HoloLens	SG (IO)	Neurosurgical applications
[119]	HoloLens	SG (NI)	Seed implantation thoracoabdminal brachytherapy
[128]	HoloLens	ST	Neurosurgical burr hole localisation
[120]	HoloLens	SG (EG)	Ureteroscopy
[105]	Moverio BT-200	SG (SSN)	Free fibula flap
[129]	Moverio BT-300	SG (EG)	Percutaneous endoscopic lumbar discectomy
[154]	HoloLens	ST	N/A
[130]	HoloLens	SG(CI)	External ventricular drainage (EVD)
[136]	HoloLens	PS	Endovascular procedures
[137]	Arzyon headset	ST	Central venous catheterisation
[106]	HoloLens	PS	Repair of complex paediatric elbow fractures
[141]	HoloLens	PS	Complex surgical procedures
[148]	HoloLens	ST	No direct surgical procedure (teaching of dental anatomy)
[169]	HoloLens	PS	Nephron-Sparing Surgery in Wilms' Tumor Surgery
[165]	HoloLens	ST	Urologic surgical procedures (bladder catheter placement)
[142]	HoloLens	SG (DG)	Right colectomy with extended lymphadenectomy
[112]	PicoLinker glasses	SG (SI)	Single-segment posterior lumbar interbody fusion
[307]	HoloLens	SG (NI)	Transjugular intrahepatic portosystemic shunt (TIPS)
[170]	HoloLens	SG (NI)	Percutaneous needle interventions
[144]	HoloLens	PS	Example use cases: laparoscopic liver resection and congenital heart surgery
[121]	HoloLens	SG (DG), PS, TELC, ST	Laparoscopic partial nephrectomy / Laparoscopic radical nephrectomy
[107]	HoloLens	SG (NI)	Percutaneous image-guided spine procedures
[108]	HoloLens	SG (DTG)	Total shoulder arthroplasty
[65]	HoloLens	SAA	Open Abdomen Surgery
[109]	HoloLens	SG (SNN)	Hallux Valgus correction
[110]	HoloLens	SG (SI)	Spinal instrumentation
[111]	HoloLens	SG (KWI)	Reverse total shoulder arthroplasty (RSA)
[171]	Metavision Meta 2	SG (EG)	Laparoscopic procedures
[143]	HoloLens	SAA	N/A
[153]	HoloLens	TELM	Cricothyroidotomy
[155]	HoloLens	SG (IO)	Surgery of the parotid gland
[156]	HoloLens	SG (IO)	Lateral Skull Base Surgery
[159]	HoloLens	SG (DG)	Perforator flap transfer
[162]	HoloLens	SG (IO)	Mandibular reconstruction

Table A3: Addressed and Persistent Human Factors (notation: human factor(s) on the left side of the arrow trigger other human factor(s) on the right side of the arrow).

Study	Addressed Human Factors	Reported Persistent Human Factors
[113]	MM, INTPN_2D_DETAIL, HEC, SLC	N/A
[166]	ATTN_SHIFT → HEC, TOOL_ADJUST, DPPC, EXP_OUTCOME, SPATIAL_PERC	VIS_OPT
[89]	ATTN_SHIFT → INC	N/A
[90]	[ATTN_SHIFT → HEC, DIST, CONC_LS], SLC	N/A
[96]	ATTN_SHIFT → HEC	EASE_HCI, SLC, VIS_OPT, COMF
[97]	ATTN_SHIFT → DIST, MM, SLC	DPPC
[98]	SUBJ_MEAS_OUTCOME, SLC	PER_REAL_AUG, FAT, IMMR, EASE_HCI, SPATIAL_PERC
[91]	ATTN_SHIFT → DIST	EYE
[99]	ATTN_SHIFT → HEC	VIS_OPT
[122]	ATTN_SHIFT, HEC	CONC_LS, ANX
[100]	ATTN_SHIFT, SLC	N/A
[101]	[INTPN_2D_DETAIL → SURG]	PER_REAL_AUG
[102]	INTRA_OP_NAV	N/A
[95]	[ATTN_SHIFT → SURG_ERR], [MM → SLC]	N/A, DPPC
[103]	[ATTN_SHIFT → HEC], [MM → FRUS], SPATIAL_PERC	N/A
[138]	ATTN_SHIFT, MM	N/A
[146]	[ATTN_SHIFT → SLC & SURG_ERR], DPPC	N/A
[157]	ATTN_SHIFT, SUBJ_MEAS_OUTCOME	PREF_HOL, PER_REAL_AUG
[158]	[INTPN_2D_DETAIL → SURG_ERR], DPPC	N/A
[151]	DPPC, COMM_3D, MM	EASE_HCI
[149]	EMP_EST_2D, EASE_HCI, CLIN_EXP_2D	EASE_HCI
[150]	EMP_EST_2D, EASE_HCI	EASE_HCI
[152]	[DPPC, INTPN_2D_DETAIL → INTRA_OP_NAV]	EASE_HCI, VIS_OPT
[160]	[ATTN_SHIFT → HEC], DPPC, SPATIAL_PERC	COMF
[131]	[ATTN_SHIFT → HEC]	N/A
[132]	MM	EASE_HCI
[123]	COMM_3D	EASE_HCI
[124]	MM, ATTN_SHIFT	SPATIAL_PERC
[167]	SPATIAL_PERC, TOOL_ADJUST, SLC	PER_REAL_AUG
[92]	[ATTN_SHIFT → SURG_ERR]	N/A
[71]	STRESS, MIP, [ATTN_SHIFT → ERG, SURG_ERR]	VIS_OPT
[163]	N/A	USEF
[94]	MM, INTRA_OP_NAV	SPATIAL_PERC, SUBJ_MEAS_OUTCOME
[66]	MM, [ATTN_SHIFT → HEC], SPATIAL_PERC	COMM_3D
[87]	SLC	N/A
[115]	SLC, MM, SPATIAL_PERC	SPATIAL_PERC, FAT
[116]	ATTN_SHIFT, MM, FRUS, DPPC	FRUS
[135]	CLIN_EXP_2D, SLC, HEC, EXP_OUTCOME	ANX, COMF, CONF
[118]	MM, SPATIAL_PERC	SPATIAL_PERC, COMF
[125]	[ATTN_SHIFT → SURG]	N/A
[133]	MM, SLC	N/A
[105]	MM, ATTN_SHIFT	N/A
[139]	ATTN_SHIFT	N/A
[126]	[ATTN_SHIFT → MM, HEC], SPATIAL_PERC	COMF
[127]	ATTN_SHIFT, MM	DPPC
[119]	ATTN_SHIFT, SURG	SLC
[168]	[ANAT_PLN → SURG]	SPATIAL_PERC
[104]	MM	SPATIAL_PERC
[128]	MM, EXP_OUTCOME, SLC, SPATIAL_PERC	SPATIAL_PERC
[129]	[ATTN_SHIFT → HEC]	COMF
[93]	ATTN_SHIFT	EASE_HCI
[134]	SPATIAL_PERC	N/A
[140]	SURG	N/A
[147]	SURG, ATTN_SHIFT, HEC	DPPC
[172]	DIST, FAT	DIST

(continued on next page)

Table A3 (continued)

Study	**Addressed Human Factors**	**Reported Persistent Human Factors**
[88]	SLC	COMF
[62]	EYE, SUBJ_MEAS_OUTCOME	SUBJ_MEAS_OUTCOME
[164]	SLC, DIST	DIST, USEF, EASE_HCI
[120]	[ATTN_SHIFT, HEC → SURG, SPATIAL_PERC]	COMF, SPATIAL_PERC
[117]	SLC, ATTN_SHIFT, HEC, SPATIAL_PERC	N/A
[161]	SURG	N/A
[114]	ATTN_SHIFT	N/A
[145]	ATTN_SHIFT	N/A
[154]	ENG_MOT	ENG_MOT, SPATIAL_PERC
[130]	EXP_OUTCOME	EASE_HCI
[136]	ATTN_SHIFT, SPATIAL_PERC	SPATIAL_PERC
[137]	SLC	USEF, EASE_HCI, FRUS
[106]	SPATIAL_PERC, CONF	SPATIAL_PERC
[141]	N/A	SPATIAL_PERC
[148]	SPATIAL_PERC, CONF	EASE_HCI, COMF, USEF
[169]	DPPC, SPATIAL_PERC	USEF, CONF
[165]	N/A	EASE_HCI, CONF, SLC
[142]	N/A	CONF, SPATIAL_PERC
[112]	[ATTN_SHIFT → SURG, INC]	N/A
[307]	SURG, SPATIAL_PERC	SURG
[170]	[HEC → SURG]	FAT, INC, COMF
[144]	SPATIAL_PERC, MM	DPPC, COMF
[121]	SPATIAL_PERC, SURG	N/A
[107]	MM	EASE_HCI
[108]	ATTN_SHIFT, SURG	N/A
[65]	ATTN_SHIFT, SURG, MM, SLC	COMF, STRESS, EASE_HCI, DPPC
[109]	SLC, EXP_OUTCOME, SURG	EXP_OUTCOME
[110]	SLC, EXP_OUTCOME, SURG, ATTN_SHIFT	N/A
[111]	SURG, EXP_OUTCOME	N/A
[171]	[ATTN_SHIFT → HEC], SLC, COMF, FAT	DPPC, COMF, EASE_HCI, USEF
[143]	N/A	EASE_HCI, USEF, VIS_OPT, COMF
[153]	[ATTN_SHIFT → MM, SURG]	EASE_HCI, FRUS
[155]	N/A	EASE_HCI, COMF
[156]	SPATIAL_PERC	N/A
[159]	DPPC, MM	N/A
[162]	N/A	N/A

Table A4: Description of AR visualisation, conducted experiments and accuracy of final 91 articles used for quantitative synthesis. Acronyms: PM: Preoperative Model; II: Intraoperative image. PI: Preoperative image. IM: Intraoperative Model. IV: Intraoperative live streaming video. PV: Preoperatively recorded video. DOC: Documents. COMM: 2D plane with video communication software application (google hangouts etc.). IND: Intraoperative Numerical Data. SSE: System Setup Experiment without phanton, cadaver or patient involvement (may contain additional hardware). PE: Phantom Experiment. HCE: Human Cadaver Experiment. AE: Animal experiment. ACE: Animal Cadaver Experiment. SE: Simulator Experiment. SCE: Simulated Clinical environment experiment. PS: Patient Case Study. Abbreviations: Quan: Quantitative Study. Qual: Qualitative Study.

Study	AR visualisations	Experiments	Reported Accuracy
[89]	PM: optimal bone drill trajectory, organs, bone structures	Quan: PE: 1.) registration accuracy, 2.) surgical navigation. HCE: 3.) joint screw implantation	1.) 0.809 ± 0.05 *mm*, $1.038^\circ \pm 0.05^\circ$.
[90]	PM: 3D pelvis model incl. vessels, optimal bone drill trajectory	Quan: HCE: joint screw implantation	2.7 ± 1.2 *mm*, $3.7 \pm 1.1 mm$, $2.9^\circ \pm 1.1^\circ$
[96]	II: radiographic images	Quan: & Qual: PE: Percutaneous vertebroplasty, kyphoplasty and discectomy interventions	N/A
[97]	II: 2D X-ray images inkl. annotations, IM: guiding lines, planes & spheres, C-arm source position (cylinder)	Qual: SSE: 1.) Calibration, 2.) HMD tracking, 3.)Landmark identification, PE: 4.) K-wire guidance, 5.) Entry point localization (implantation of nails)	1.) 21.4 ± 11.4 *mm*, 2.) 16.2 ± 9.5 *mm*, 3.) $8.76 - 11.7 \pm (3.21 - 4.03)$ *mm*, 4.) 4.47 ± 2.91 \$ 9.84 ± 3.97, 5.) 5.2 *mm*
[98]	PM: Anatomical 3D models (incl. bones & muscles), virtual menu with toggle buttons, preoperative plan. IM: optimal tool trajectory	Quan: 1.) System accuracy estimation (perceived AR target positions). Qual: 2.) Subjective workload assessments (NASA Task Load Index)	1.) 0.6 *mm*
[91]	IM: pose of surgical tool (stack of cyan rings), navigation target (circle)	Quan: PE: 1.) tracked tool positioning & orienting, Qual: 2.) questionnaire	1.) 0.40 ± 0.78 *mm*, $2.07 \pm 1.68^\circ$
[99]	PM: virtual trajectories (pedicle screw guidance), lumbar spine 2D & 3D CT images	Quan: PE: 1.) Registration accuracy verification, 2.) Percutaneous placement	1.) 12.99 mm (12.31–13.61 mm), 2.) 12.99 mm (12.31–13.61 mm), 15.59 mm (12.02–18.69 mm)
[100]	PM: 3D organs incl. fiducial or anatomical markers	Quan: 1.) reliability assessment of virtual-physical mappings, Quan: & Qual: 2.) assessment of superimposed 3D virtual overlays in physical space	1.) $19.74 \pm 2.38 mm$ (x), $76.82 \pm 3.83 mm$ (y), $2.74 \pm 1.96 mm$ (z), $19.74 \pm 2.38^\circ$, 76.82 ± 3.83, $2.74^\circ \pm 1.96^\circ$, 2.) 3.2 ± 1.6 mm (RMSE)
[101]	PI: anteroposterior lumbar X-ray 2D images	Qual: PS: 1.) Accuracy and 2.) repeatability validation	1.) 8.77 mm
[102]	PM: targeted screw trajectory, drill entry points. IM: drill angle between current and targeted screw trajectory, 3D trajectory deviation triangle	Quan: PE: guiding wire placement for pedicle screw	2.77 ± 1.46 mm, $3.38^\circ \pm 1.73^\circ$
[103]	II: interventional X-ray images, IM: view frustrum	Quan: SSE: 1.) Hand-eye calibration experiment, PE: 2.) internal fixation of pelvic ring fractures & percutaneous vertebroplasty	1.) 0.43 mm ± 0.34 mm, 0.43° ± 0.34°
[113]	IM: needle visualisations (needle position, orientation & shape, tangential ray)	Quan & Qual: PE: needle insertion task	8.15 mm ± 0.4 mm, 6.54 mm ± 0.294 mm, 6.03 mm ± 0.291 mm
[138]	PM: 3D objects (cube)	Quan: PE: 1.) Tight-Fit & Loose-Fit Accuracy Evaluation	1.) 0.7 ± 0.2 mm, 2.) 2.3 ± 0.5 mm
[62]	3D calibration cubes	Quan: SSE: calibration accuracy	below 6 mm, up to 5°

(continued on next page)

Table A4 (continued)

Study	AR visualisations	Experiments	Reported Accuracy
[151]	PM: 3D heart models	Quan: SSE: patient based heart model analysis (anatomy identification & diagnosis)	N/A
[149]	PM: 3D coronay arteries models	Quan: SSE: Dynamic and static gesture recognition	N/A
[150]	PM: 3D cardio artery vascular models	Quan: SSE: hand gesture recognition rate validation	N/A
[152]	PM: 3D heart, spine & cathether models, 3D catheter path planning	Quan: PE: catheter navigation under C-arm fluoroscopy guidance	0.425 ± 0.021 mm (registration), 0.29 ± 0.19 mm (catheter position)
[131]	II: ultrasound images	Quan: SE: sonographic guided jugular vein catheterization	N/A
[132]	PM: 3D patient surface mesh, vascular tree, catheter position, registration landmarks,	Quan: PE: 1.) calibration, 2.) catheter insertion & navigation, 3.) Likert scale questionnaire evaluation	1.) 1.) 4.34 ± 0.709 mm (RMSE point-to-point correspondence)
[123]	PM: complex medical vascular & blood flow 3D image data	Quan: SSE: Evaluation of vascular & blood flow image data	N/A
[124]	PM: 3D skull visualisations, localisation markers	Quan: PE: 1.) Manual registration, 2.) Maintaining 3D virtual content registration via continuous camera tracking	1.) 4.39 ± 1.29 mm, 2.) 1.4 ± 0.67 mm (mean perceived drift)
[122]	II: 2D neuronavigation images	Quan: PS: pedicle screw placement	N/A
[160]	II: 2D navigation monitor. PM: 3D mandible model, 3D osteotomy cutting guides planes)	Quan: PE: osteotomies: 1.) augmented navigation system monitor & 2.) superimposition of surgical plan	1.) 1.79 ± 0.94 mm, $3.67 \pm 3.67°$, 2.) 2.41 ± 1.34 mm,
[157]	PM: Preoperative & ideal postoperative3D facial surface and facial bones	Quan: PS: reconstructive surgies (facial fractures or deformities)	$30-40$ mm (display error)
[158]	PM: 3D bony, vascular, skin & soft tissue structures, vascular perforators, bounding box	Quan: PS: flap surgery	N/A
[167]	PM: 3D virtual robot arm, 2D reflective AR display	Quan: PE: Registration with 1.) and without 2.) reflective AR displays, 2.) Simulated robot-assisted trocar placement	1.) 16.5 ± 11.0 mm, 2.) 30.2 ± 23.9 mm (misalignment error)
[166]	PM: & II: 3D plane with endoscopy visualization, IM: viewing frustrum, PM: ndoscope,	Quan: SSE: 1.) Display calibration, 2.) Camera calibration. PE: Visualization performance evaluation	1.) 4.27 ± 3.09 mm
[146]	PI: 2D radiographic images with guidance information, IM: 3D drill guidance nformation	Quan: PE: 1.) Accuracy evaluation, 2.) Tool navigation & guidance	1.) Avg: 0.46 mm, Max: 0.86 mm, Avg: 1.17° , Max: 2.10°
[92]	II: fluoroscopic video	Quan: PE: Guide wire insertion into femur	2.6 ± 0.02 mm
[71]	PM: & IM: position, depth & alignment of planed & actual dental drill, injury avoidance warnings, drill heads	Quan: 1.) SSE: Calibration accuracy, 2.) ACE: Implant placement	1.) 3.01 ± 3.01 mm, 2.) < 2.5 mm (implant deviation),
[163]	II: preoperative CT scan	Quan: PS: different urological procedures, Likert scale questionnaire	N/A
[94]	IM: Live 3D point cloud (C-arm pose)	Quan: PE: pelvic trauma surgery	51.6 ± 19.2 mm, $1.54 \pm 0.92°$
[66]	PM: 3D hepatic artery, portal vein, hepatic veins, liver tumor, liver capsule	Quan: PS: open hepatic surgery	N/A
[87]	COMM: google hangouts, DOC: articles from senior author	Quan: PS: reconstructive limb salvage procedure	N/A
[115]	PM: 3D anatomical models, 3D ultrasound streaming plane	Quan: SE: transesophageal echocardiography	N/A
[116]	PM: 3D graphical annotations (incision lines, surgical instruments)	Quan: SE: 1.) anatomical marker placement, 2.) mock abdominal incision	1.) 11.37 ± 0.72 mm

(continued on next page)

Table A4 (continued)

Study	AR visualisations	Experiments	Reported Accuracy
[117]	PM: 3D liver structure (intraoperatively updated), tumor, virtual needle, egistration	Quan: 1.) PE: Registration accuracy validation, 2.) AE: needle insertion operation	1.) 2.24 mm (avg. target registration error)
[135]	PM: 3D graphical annotations (lines & models)	Quan: HCE: leg fasciotomy	N/A.
[118]	PM: 3D liver incl. parenchyma, portal, hepatic veins & lesion	Quan: SSE: Identification of liver segments	N/A
[161]	PM: 3D landmarks, 3D tumors, 3D axial facial CT slice	Quan: PE: Automatic registration after user calibration	x: 3.3 ± 2.3 mm y: -4.5 ± 2.9 mm z: -9.3 ± 6.1 mm
[125]	PM: 3D patient head incl. skin, skull & spine	Quan: PE: Registration accuracy. 3 registration methods: 1.) Keyboard, 2.) Tap to Place, 3.) 3-Point correspondence matching	1.) X Axis: 5 ± 5° Y Axis: -5.9 ± 5.9° Z Axis: 6.8 ± 5.9°; displacement: XY Plane: 2.9 ± 1.8 mm ZY Plane: 1.8 ± 1.2 mm XZ Plane: 1.6 ± 0.9 mm
[133]	PM: 3D organs, needle (actual & preoperative plan)	Quan: 1.) PE: & 2.) AE: needle insertion	1.) 0.664 mm, 4.74°, 2.) 1.617 mm, 5.574°
[105]	PM: 3D bones, surgical plan: control points, osteotomy trajectories, navigated saw, 2D digital coordinate system	Quan: PE: osteotomy	4.1 ± 2.29 mm, 5.08 ± 3.64°, 4.97 ± 2.91°
[139]	PM: 3D patient skin surface	Quan: PE: Alignment (different data sparsity percentages are tested but we refer only to 100 % of floating data being used)	5 reference points alignment error RMSE: Avg.: 0.932 mm, Min: 0.37 mm, Max: 1.49 mm
[126]	PM: 3D intracranial structure, lesion	Quan: PS: Craniotomy	N/A.
[127]	PM: 3D Needle insertion guidance visualization options: 1.) planes, 2.) lines, 3.) cone rings	Quan: PE: 1.) Registration accuracy estimation: a) Angle measurement of displayed lines I, b) Angle measurement of displayed lines II, c) Tracked normal vector accuracy, d) Tracked normal vector accuracy, 2.) Comparison study	1 a.) 0.76 ± 0.11°, 1 b.) frontal viewing pos. 1.90 ± 1.82°, 45° viewing pos. 4.28 ± 4.09°, lateral viewing pos. 7.94 ± 7.75°, 1 c.) 0.72 ± 0.41°, 1 d.) X $ Y marker: 0.27 ± 0.21°, X $ Z marker: 0.31 ± 0.22°, Y $ Z marker: 0.38 ± 0.36°
[119]	3D anatomy (skin, bones, tumor tissue), virtual needles (planning & detected), seeds, 2D control panel	Quan: 1.) PE: & 2.) AE: brachytherapy of tumors	Avg. needle location error: 1.) 0.957 mm, 2.) 2.416 mm
[168]	PM: 3D kidneys incl. tumor, arteries, veins, urinary collecting structures	Quan: SSE: Assessment of anatomical structures	N/A
[104]	IM: 3D anatomical structures, C-arm principle axis	Quan: SSE: 1.) calibration accuracy, PE: 2.) Target augmentation error, 3.) Augmented surgical visualisation	1.) 5.7 ± 0.26 mm, 2.) 10.8 ± 3.45 mm
[128]	PM: 3D patient skin surface, brain, intra-cortical lesion	Quan: PE: Target Localisation	N/A
[129]	II: live endoscopic camera image	Quan: PS: Lumbar discectomy	N/A
[93]	COMM: videoconferencing application, PI: patient records	Quan: PE: SCE: Mobile access to patient records, telepresence	N/A
[134]	PM: 3D arterial system, aneurysm, bones, PI: 2D image with volume rendering, arterial diameters & planning notes	Quan: PS: Abdominal aortic aneurysm repair	N/A
[140]	PM: 2D surgical safety checklist	Quan: SSE: time-out checklist execution	N/A
[147]	PM: 3D tooth, cone (endoscope view frustrum), probe alignment cyclinder & planes, II: 2D imaging	Quan: PE: 1.) Augmentation quality evaluation, 2.) Dental decay localization	31 ± 11 px (keypoint displacement)
[172]	IND: 2D screen incl. patient heart rate, blood pressure, blood oxygen saturation, alarm notifications	Quan: & Qual: PS: vital sign monitoring, Quan: situation awareness measurement	N/A
[88]	IV: hybrid image (surgical field combined with hands of remote surgeon)	Quan: PS: shoulder replacement	N/A

(continued on next page)

Table A4 (continued)

Study	AR visualisations	Experiments	Reported Accuracy
[164]	IV: interactive video display incl. cursor moved by supervising physician, PV: training guide	Qual: & Quan: SSE: user survey	N/A
[120]	PI: CT images, IV: live fluoroscopy, endoscopic view	Qual: & Quan: SE: mid-ureteric stone removal	N/A
[114]	PM: 3D patient anatomy (e.g. head, intracranial vascular tissue)	Quan: PE: dummy head alignment test	< 3 mm (Avg. Target Registration Error)
[95]	PM: 3D bone structures, fiducial markers	Quan: PE: accuracy assessment	Fiducual marker comparisons (RMSE): x: 3.22 mm, y: 22.46 mm, z: 28.30 mm
[145]	IND: 2D screen incl. patient arterial line blood pressure, heart rate, heart rhythm, pulse oximetry, respiratory rate	Quan: SE: Vital signs monitoring during bronchoscopy	N/A
[154]	PM: 3D ear anatomy	Quan: SSE: spatial exploration of 3D virtual ear model	N/A
[130]	PM: 3D catheter	Quan: SSE: 1.) Stability measurement of tracking algorithm, 2.) testing of tracking accuracy 3.) latency test using third-party tracker, 4.) HCE: EDV performed on a cadaveric head	2.) avg. distance from catheter tip to corresponding grid intersections (2D plane): 0.58 mm, overall avg. accuracy on all 3 grid faces: 0.85 mm (3D space)
[136]	PM: 3D volumes from MRI images	Quan: AE: transarterial embolization of hepatocellular carcinoma (HCC)	N/A
[137]	PM: 3D model of body simulator's external surface (upper torso) and 3D vascular structures	Quan: SE: tracked needle insertion	N/A
[106]	PM: 3D elbow fractures (bones)	Quan: SSE: Orthopedic surgeons' assessment of 3D AR models for presurgical planning in complex pediatric elbow fractures	N/A
[141]	PM: 3D spine model with a vascular model overlay	Quan: SSE: 3D model measurement using circumference and angle tools of standard-of-care PACS software	N/A
[148]	PM: 3D human skull	Quan: SSE: digital anatomy session with the HoloHuman virtual anatomy training software	N/A
[169]	PM: 3D intraparenchymal arteries and veins, kidney, tumor	Quan: PS: Preoperative planning of patients eligible for nephron-sparing surgery (NSS)	N/A
[165]	PV: 2D plane with catheter placement instruction guidance	Quan: SE: Bladder catheter placement using a male catheterization-training model	N/A
[142]	PM: 3D anatomy models (e.g. vascular model)	Quan: PE: registration and needle placement	Target error distance: x-axis: 2.9757 ± 1.33396 mm, y-axis: 2.2790 ± 1.44992 mm, z-axis: $2.7844 \pm .91323$mm
[112]	II: fluoroscopic 2D image	Quan: PS: single-segment posterior lumbar interbody fusion (PLIF) at L5–S1	N/A
[307]	PM: 3D portal vein and hepatic vein, liver	Quan: 1.) AE: dogs: simulated percutaneous puncture of the portal vein and simulated TIPS, 2.) PE: liver phantom experiment	N/A
[170]	PM: 3D internal organs, PI: CT images, IM: progress view of the virtual planned target, needle path, skin entry point and needle end	Quan: PE: 1.) image overlay accuracy using 3D abdominal phantom, 2.) needle placement performance using tissue phantom	1.) Total target overlay error over 336 targets: 1.74 ± 0.86 mm. Needle overlay angle: $0.41 \pm 0.23°$
[144]	PM: 3D liver and heart, slicing tool (plane)	SSE: visualisation of patient-specific models and interaction with 3D virtual objects (rotate, scale and move)	N/A

(continued on next page)

Table A4 (continued)

Study	AR visualisations	Experiments	Reported Accuracy
[121]	PM: 3D target organs and tumors (kidney, tumor, renal vessels, renal collection system, skin, skeleton, liver, spleen), PI: MR results, IV: laparoscopic video stream	Quan: PS: Prospective review of patients with stage T1N0M0 renal tumors who untervent laparoscopic partial ephrectomy	N/A
[107]	PI: 3D plane with axial CT image, PM: needle trajectories in correct spatial orientation over patient	Quan: 1.) PE: control data experiment (needle navigation) using skull with ballistic gelatin and radiopaque balls (targets), 2.) PS: interventional spine procedures	1.) 0.998 ± 1.66 mm (mean error of needle tip to targeted ball). Mean distance from model surface to targeted ball: 79.42 ± 15.33 mm, 2.) Mean error of needle to target: 1.73 ± 2.20 mm
[108]	PM: 3D glenoid and planed drilling path	Quan: PE: 1.) inside-out registration (via HoloLens depth sensing camera), 2.) accuracy evaluation of inside-out registration using outside-in tracking with optical tracker, 3.) registration with surface digitisation	1.) Inside-out registration accuracy compared with external tracking (optical tracker is used to verify inside-out tracking): translation (max: 21.82 ± 2.33 mm), rotation: max. 8.10 ± 2.89°
[65]	PM: 3D patient anatomy	Quan: PS: open abdomen surgeries	N/A
[109]	PM: 3D foot	Quan: PE: distal osteotomy	Mean deviation between osteotomy plane and target plane perpendicular to the second metatarsal (anterior direction): 1.) Experienced surgeons: 4.9 ± 4.2°, 2.) less experienced surgeons: 6.4 ± 3.5°
[110]	PM: vertebral body	Quan: PE: drilling pilot holes in lumbar vertebra sawbones models	average minimal distance of the drill axis to the pedicle wall (MAPW): 1.) Expert surgeons: 5.0 ± 1.4 mm, novice surgeons: 4.2 ± 1.8 mm
[111]	PM: planned drill trajectory, IM: current drill trajectory, deviation in degrees and millimeters	Quan: PE: Using 3D printed scapula based on scans of human cadavers: Guidewire positioning of the central back of he	mean deviation of placed guidewires from the planned trajectory: 2.7 ± 1.3°, mean deviation to the planned entry point of the placed guidewires: 2.3 ± 1.1 mm
[171]	IV: 2D plane with laparoscopic video feed, PI: 2D plane with MRI image sclices	Qual: & Quan: SE: laparoscopic training simulator incl. MRI images and pre-recorded laparoscopic video feed	N/A
[143]	PM: 3D organ models (brain and liver), PI: 2D planes with volumetric /CT data (with scrolling bar), 2D plane with intraoperative data (pCLE, iUS) (with transparency adjustment up and down arrows))	Quan: SSE: interaction with the visualisation components and exploration of 3D virtual functionalities	N/A
[153]	PM: 3D annotations (incision lines) and 3D surgical tools	Quan: SE: performing cricothyroidotomies in a simulated austere scenario (smoke and loud noises of gunshots and explosions)	N/A
[155]	PM: 3D mandible, parotid, tumor, head, grey circles, operating menu (buttons), PI: 2D MRI images	Quan: PS: live parotid surgery: study persons who did not participate in the actual surgery performed manual ologram	Manual Registration accuracy using fiduical markers on the head: outer borders of face: 10.09 ± 4.23mm, arotid: 13.39 ± 4.71 mm, Tumor: 13.29 ± 5.66 mm
[156]	PM: 3D skull and temporal bone	Quan: PE: Evaluation of manual target registration error using skull model	Target registration error: 10.62 ± 5.90 mm10.62
[159]	PM: 3D vascular map, surrounding soft tissues, marker	Quan: PE: Precision verification of the vascular localisation system	mean errors (under different conditions): min: 1.35 ± 0.43 mm , max: 3.18 ± 1.32
[162]	PM: 3D planned mandibular reconstruction result	Quan: PE: 1.) Accuracy validation experiment for OST-HMD calibration (3D printed skull with fiducials), 2.) Calibration method testing, 3.) PS: mandibular reconstruction	1.) Avg. root-mean-square error of control points between rendered object and skull model: 1.30 ± 0.39

Appendix B

Implementation Details of the Synthetic Image Dataset Generation Pipeline in Chapter 4

This appendix provides supplementary material intended to support the reader in understanding key aspects of the synthetic image generation method used for markerless 3D hand and tool pose estimation, as presented in Chapter 4. Section B.1 describes the camera sphere parameters required to programmatically generate the virtual sphere around the hand holding the tool. Section B.2 outlines the parameters used to control grasp generation and rendering within the proposed pipeline.

B.1 Explanation of Camera Sphere Parameters

To support a clearer understanding of the mathematical concepts behind the sphere-based camera view generation method defined in Equations 4.1 and 4.2, Chapter 4, this appendix provides a detailed explanation of the relevant parameters. Firstly, it outlines the parameters of the number of latitude segments and circles per segment. Secondly, it describes the spherical coordinates used to define the camera viewpoints.

Number of Latitude Segments (N^{ϕ}) and Circles per Segment ($N_{circ}^{(i)}$)

- r_{circ}: The angular (geodesic) radius of each viewing circle (spherical cap) on the sphere. For a given sphere configuration, all circles use the same r_{circ}. This parameter controls the coverage–overlap trade–off: larger $r_{circ} \Rightarrow$ more overlap, fewer gaps.

- r_{sph}: The radius of the sphere on which the camera viewpoints are distributed, i.e. the overall size of the area being considered for viewpoint placement.

- N^{ϕ}: The number of latitude floors on the sphere, which is calculated based on the ratio of the circle radius (r_{circ}) to the sphere radius (r_{sph}). N^{ϕ} is used to divide the sphere into horizontal segments, where each segment's height is determined in such a way that it roughly matches the diameters of the circles ($2r_{circ}$).

- $\lfloor \cdot \rfloor$: The floor function, which rounds down to the nearest integer, and ensures that the number of segments and circles per segment are whole numbers.

- $N_{circ}^{(i)}$: The number of circles in the i-th latitude segment, based on the circumference of the sphere at latitude θ_i divided by the diameter of the circles, ensuring even spacing of the viewpoints.

- θ_i: The colatitude angle for the i-th latitude segment, with $\theta_i \in \{0, \ldots, \pi\}$

Spherical Camera Viewpoint Coordinates

- $(\theta_i, \phi_j^{(i)})$: The spherical coordinates that define the camera viewpoints. θ_i is defined above. $\phi_j^{(i)} = j \cdot \frac{2\pi}{N_{circ}^{(i)}}$ is the azimuthal angle for the j-th circle within the i-th latitude segment. The term $\frac{2\pi}{N_{circ}^{(i)}}$ ensures an even distribution around the latitude circle.

B.2 Grasp Generation and Rendering Pipeline Configuration

Table B1 outlines the different parameters of the frame generation pipeline, which influence the diversity and quantity of frames produced. Using the variable notations of Table B1, the frame generation used to create the dataset can be expressed as in Algorithm 1. This algorithmic formulation highlights the combinatorial structure of the pipeline and its potential for scalable dataset generation.

Algorithm 1 Combinatorics for frame generation based on a selected grasp

```
Input: Selected grasp $(G_{[\Theta_{\mathrm{Vol}}]})_k$, egocentric viewpoint $z_{\mathrm{ego}}$, background image $\mathscr{G}_{\mathrm{bkgr}}$, hand texture $H_{\mathrm{rgb}}$, camera Euler angles $\Theta_k$, sphere radius $r_{\mathrm{sph}}$, circle radius $r_{\mathrm{circ}}$
Output: Set of frames $F$
Initialize $F \leftarrow \emptyset$
for all $(G_{[\Theta_{\mathrm{Vol}}]})_k, z_{\mathrm{ego}}, \mathscr{G}_{\mathrm{bkgr}}, H_{\mathrm{rgb}}, \Theta_k$ do
    $frame \leftarrow$ GenerateFrame($(G_{[\Theta_{\mathrm{Vol}}]})_k, z_{\mathrm{ego}}, \mathscr{G}_{\mathrm{bkgr}}, H_{\mathrm{rgb}}, \Theta_k, r_{\mathrm{sph}}, r_{\mathrm{circ}}$)
    $F \leftarrow F \cup \{frame\}$
end for
```

Table B1: Frame generation configuration

Parameter	**Description**	**Values**
$(\Theta_{Vol})_j$	k-th ($k \in [1,..,N]$) probe rotation Euler angles, used within GrabNet	Manually selected set of three Euler angles $[\alpha,\beta,\gamma \vert \alpha,\beta,\gamma \in [0,...,360]]$
$(G_{[\Theta_{Vol}]})_k$	k-th manually selected GrabNet-based grasp out of all grasps generated by Θ_{Vol}	11 manually selected plausible grasps via visual inspection in MeshLab
r_{sph}	Sphere radius (egocentric camera distance; eye–to–hand)	$\{0.5, 0.8\}$ m *Note:* Both values of r_{sph} are used to render a full set of viewpoints over the sphere at each eye–to–hand distance.
r_{circ}	Sphere surface circle radius	0.15 m
Θ_k	k-th ($k \in [1,..,92]$) camera view point Euler angles that depend on r_{sph} and r_{circ}	92 - 2 excluded = 90 values. Concrete Euler angles can be derived from applying the sphere concept described in Section] 4.2.3 of Chaper 4
$\mathscr{G}_{bkgr}$	Background image for rendered grasp frames	1 x plain white, 1 x consultation room, 3 x SPACE-FAN phantom, 3 x real pregnant mother belly (white / brown / black)
H_{rgb}	RGB values of glove texture	$\begin{bmatrix} 0.5647058824 \\ 0.5921568627 \\ 0.768627451 \end{bmatrix},$ $\begin{bmatrix} 0.38039215686 \\ 0.61960784314 \\ 0.8666666667 \end{bmatrix}$
A_{rgb}	RGB values of the arm	$[1.0, 0.6784313725, 0.3764705882]$

Bibliography

[1] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments/MIT press*, 1997.

[2] Carlos E Mendoza-Ramírez, Juan C Tudon-Martinez, Luis C Félix-Herrán, Jorge de J Lozoya-Santos, and Adriana Vargas-Martínez. Augmented reality: survey. *Applied Sciences*, 13(18):10491, 2023.

[3] Ivan E Sutherland. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pages 757–764, 1968.

[4] Jiaqi Xu and Fernando Moreu. A review of augmented reality applications in civil infrastructure during the 4th industrial revolution. *Frontiers in Built Environment*, 7:640732, 2021.

[5] Ginés Morales Méndez and Francisco del Cerro Velázquez. Augmented reality in industry 4.0 assistance and training areas: A systematic literature review and bibliometric analysis. *Electronics*, 13(6):1147, 2024.

[6] Andrea Lastrucci, Yannick Wandael, Angelo Barra, Renzo Ricci, Giovanni Maccioni, Antonia Pirrera, and Daniele Giansanti. Exploring augmented reality integration in diagnostic imaging: Myth or reality? *Diagnostics*, 14(13):1333, 2024.

[7] Hitesh Chopra, Kavita Munjal, Sonia Arora, Shabana Bibi, and Partha Biswas. Role of augmented reality in surgery. *International Journal of Surgery*, 110(5):2526–2528, 2024.

[8] Poshmaal Dhar, Tetyana Rocks, Rasika M Samarasinghe, Garth Stephenson, and Craig Smith. Augmented reality in medical education: students' experiences and learning outcomes. *Medical education online*, 26(1):1953953, 2021.

[9] Fabrizio Cutolo, Nadia Cattari, Umberto Fontana, and Vincenzo Ferrari. Optical see-through head-mounted displays with short focal distance: Conditions for mitigating parallax-related registration error. *Frontiers in Robotics and AI*, 7:572001, 2020.

[10] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994.

[11] Tam Le Phuc Do, Kang Sanhae, Leehwan Hwang, and Seunghyun Lee. Real-time spatial mapping in architectural visualization: A comparison among mixed reality devices. *Sensors*, 24(14):4727, 2024.

[12] Nick T de Groot, Jasper M van der Zee, Guus MJ Bökkerink, Annemieke S Littooij, Caroline CC Hulsker, Cecilia EJ Terwisscha van Scheltinga, Cornelis P van de Ven, Ruud C Wortel, Aart J Klijn, Marc HWA Wijnen, et al. Introducing holographic surgical navigation in pediatric wilms' tumor patients: A feasibility study during total nephrectomy. *Bioengineering*, 12(8):896, 2025.

[13] Gabriel Pokorny, Rodrigo Amaral, Fernando Marcelino, Rafael Moriguchi, Igor Barreira, Marcelo Yozo, and Luiz Pimenta. Minimally invasive versus open surgery for degenerative lumbar pathologies: a systematic review and meta-analysis. *European Spine Journal*, 31(10):2502–2526, 2022.

[14] Giorgio Gandaglia, Khurshid R. Ghani, Akshay Sood, Jessica R. Meyers, Jesse D. Sammon, Marianne Schmid, Briony Varda, Alberto Briganti, Francesco Montorsi, Maxine Sun, Mani Menon, Adam S. Kibel, and Quoc-Dien Trinh. Effect of minimally invasive surgery on the risk for surgical site

infections: Results from the national surgical quality improvement program (nsqip) database. *JAMA Surgery*, 149(10):1039–1044, 10 2014.

[15] Stephanie F. Sweitzer, Emily E. Sickbert-Bennett, Jessica Seidelman, Deverick J. Anderson, Moe R. Lim, and David J. Weber. The impact of minimally invasive surgical approaches on surgical-site infections. *Infection Control 38; Hospital Epidemiology*, 45(5):557–561, 2024.

[16] Tim Xu, Susan M. Hutfless, Michol A. Cooper, Mo Zhou, Allan B. Massie, and Martin A. Makary. Hospital cost implications of increased use of minimally invasive surgery. *JAMA Surgery*, 150(5):489–490, 05 2015.

[17] Patrick J. Kelly, Jr. Alker, George J., and Stephan Goerss. Computer-assisted Stereotactic Laser Microsurgery for the Treatment of Intracranial Neoplasms. *Neurosurgery*, 10(3):324–331, 1982.

[18] David W. Roberts, John W. Strohbehn, John F. Hatch, William Murray, and Hans Kettenberger. A frameless stereotaxic integration of computerized tomographic imaging and the operating microscope. *Journal of Neurosurgery*, 65(4):545 – 549, 1986.

[19] Eleonora Barcali, Ernesto Iadanza, Leonardo Manetti, Piergiorgio Francia, Cosimo Nardi, and Leonardo Bocchi. Augmented reality in surgery: a scoping review. *Applied Sciences*, 12(14):6890, 2022.

[20] Sandro F Fucentese and Peter P Koch. A novel augmented reality-based surgical guidance system for total knee arthroplasty. *Archives of Orthopaedic and Trauma Surgery*, pages 1–7, 2021.

[21] Hasan Sumdani, Pedro Aguilar-Salinas, Mauricio J Avila, Samuel R Barber, and Travis Dumont. Utility of augmented reality and virtual reality in spine surgery: a systematic review of the literature. *World neurosurgery*, 161:e8–e17, 2022.

[22] Shivali Malhotra, Osama Halabi, Sarada Prasad Dakua, Jhasketan Padhan, Santu Paul, and Waseem Palliyali. Augmented reality in surgical navigation: a review of evaluation and validation metrics. *Applied Sciences*, 13(3):1629, 2023.

[23] Delia Cannizzaro, Ismail Zaed, Adrian Safa, Alice JM Jelmoni, Antonio Composto, Andrea Bisoglio, Kyra Schmeizer, Ana C Becker, Andrea Pizzi, Andrea Cardia, et al. Augmented reality in neurosurgery, state of art and future projections. a systematic review. *Frontiers in surgery*, 9:864792, 2022.

[24] T. Sielhorst, M. Feuerstein, and N. Navab. Advanced medical displays: A literature review of augmented reality. *Journal of Display Technology*, 4(4):451–467, 2008.

[25] Yoones A Sekhavat and Mohammad S Namani. Projection-based ar: Effective visual feedback in gait rehabilitation. *IEEE Transactions on Human-Machine Systems*, 48(6):626–636, 2018.

[26] Maria Jesus Vinolo Gil, Gloria Gonzalez-Medina, David Lucena-Anton, Veronica Perez-Cabezas, María Del Carmen Ruiz-Molinero, and Rocío Martín-Valero. Augmented reality in physical therapy: systematic review and meta-analysis. *JMIR Serious Games*, 9(4):e30985, 2021.

[27] Robert Krempien, Harald Hoppe, Lüder Kahrs, Sascha Daeuber, Oliver Schorr, Georg Eggers, Marc Bischof, Marc W Munter, Juergen Debus, and Wolfgang Harms. Projector-based augmented reality for intuitive intraoperative guidance in image-guided 3d interstitial brachytherapy. *International Journal of Radiation Oncology* Biology* Physics*, 70(3):944–952, 2008.

[28] Nassir Navab, Alejandro Martin-Gomez, Matthias Seibold, Michael Sommersperger, Tianyu Song, Alexander Winkler, Kevin Yu, and Ulrich Eck. Medical augmented reality: definition, principle components, domain modeling, and design-development-validation process. *Journal of Imaging*, 9(1):4, 2022.

[29] Michael Bajura, Henry Fuchs, and Ryutarou Ohbuchi. Merging virtual objects with the real world: Seeing ultrasound imagery within the patient. *ACM SIGGRAPH Computer Graphics*, 26(2):203–210, 1992.

[30] A. State, D.T. Chen, C. Tector, A. Brandt, Hong Chen, R. Ohbuchi, M. Bajura, and H. Fuchs. Observing a volume rendered fetus within a pregnant patient. In *Proceedings Visualization '94*, pages 364–368, 1994.

[31] W Freysinger, AR Gunkel, and WF Thumfart. Image-guided endoscopic ent surgery. *European archives of oto-rhino-laryngology*, 254:343–346, 1997.

[32] Richard Holloway. An analysis of registration errors in a see-through head-mounted display system for craniofacial surgery planning. *Unpublished doctoral dissertation, University of North Carolina at Chapel Hill*, 1994.

[33] Wolfgang Birkfellner, Klaus Huber, Franz Watzinger, Michael Figl, Felix Wanschitz, Rudolf Hanel, Dietmar Rafolt, Rolf Ewers, and Helmar Bergmann. Development of the varioscope ar. a see-through hmd for computer-aided surgery. In *Proceedings IEEE and ACM International Symposium on Augmented Reality (ISAR 2000)*, pages 54–59. IEEE, 2000.

[34] Wolfgang Birkfellner, Michael Figl, Klaus Huber, Franz Watzinger, Felix Wanschitz, Johann Hummel, Rudolf Hanel, Wolfgang Greimel, Peter Homolka, Rolf Ewers, et al. A head-mounted operating binocular for augmented reality visualization in medicine-design and initial evaluation. *IEEE transactions on medical imaging*, 21(8):991–997, 2002.

[35] Michael Rosenthal, Andrei State, Joohi Lee, Gentaro Hirota, Jeremy Ackerman, Kurtis Keller, Etta D Pisano, Michael Jiroutek, Keith Muller, and Henry Fuchs. Augmented reality guidance for needle biopsies: an initial randomized, controlled trial in phantoms. *Medical Image Analysis*, 6(3):313–320, 2002.

[36] G Goebbels, K Troche, M Braun, A Ivanovic, A Grab, K von Löbtow, HF Zeilhofer, R Sader, F Thieringer, K Albrecht, et al. Development of

an augmented reality system for intra-operative navigation in maxillo-facial surgery. *Proceedings AR/VR-Statustagung, Leipzig*, 2004.

[37] Terry M Peters, Cristian A Linte, Ziv Yaniv, and Jacqueline Williams. *Mixed and augmented reality in medicine*. CRC Press, 2018.

[38] Ryan Beams, Ellenor Brown, Wei-Chung Cheng, Janell S Joyner, Andrea S Kim, Kimberly Kontson, Dimitri Amiras, Tassilo Baeuerle, Walter Greenleaf, Rafael J Grossmann, et al. Evaluation challenges for the application of extended reality devices in medicine. *Journal of Digital Imaging*, 35(5):1409–1418, 2022.

[39] Longfei Ma, Tianqi Huang, Jie Wang, and Hongen Liao. Visualization, registration and tracking techniques for augmented reality guided surgery: a review. *Physics in Medicine & Biology*, 68(4):04TR02, 2023.

[40] Sheng-Xian Xiao, Wen-Tien Wu, Tzai-Chiu Yu, Ing-Ho Chen, and Kuang-Ting Yeh. Augmenting reality in spinal surgery: A narrative review of augmented reality applications in pedicle screw instrumentation. *Medicina*, 60(9):1485, 2024.

[41] Arrigo Palumbo. Microsoft hololens 2 in medical and healthcare context: state of the art and future prospects. *Sensors*, 22(20):7709, 2022.

[42] S. Condino, M. Carbone, R. Piazza, M. Ferrari, and V. Ferrari. Perceptual limits of optical see-through visors for augmented reality guidance of manual tasks. *IEEE Transactions on Biomedical Engineering*, 67(2):411–419, 2020.

[43] David M Hoffman, Ahna R Girshick, Kurt Akeley, and Martin S Banks. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, 8(3):33–33, 2008.

[44] Mitchell Doughty and Nilesh R Ghugre. Head-mounted display-based augmented reality for image-guided media delivery to the heart: a preliminary investigation of perceptual accuracy. *Journal of Imaging*, 8(2):33, 2022.

[45] Anil Ufuk Batmaz, Mayra Donaji Barrera Machuca, Junwei Sun, and Wolfgang Stuerzlinger. The effect of the vergence-accommodation conflict on virtual hand pointing in immersive displays. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.

[46] Jannick P Rolland and Henry Fuchs. Optical versus video see-through head-mounted displays in medical visualization. *Presence: Teleoperators & Virtual Environments*, 9(3):287–309, 2000.

[47] Mitchell Doughty, Nilesh R. Ghugre, and Graham A. Wright. Augmenting performance: A systematic review of optical see-through head-mounted displays in surgery. *Journal of Imaging*, 8(7), 2022.

[48] Ruiyang Li, Boxuan Han, Haowei Li, Longfei Ma, Xinran Zhang, Zhe Zhao, and Hongen Liao. A comparative evaluation of optical see-through augmented reality in surgical guidance. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[49] Jian Ye, Qingwen Chen, Tao Zhong, Jian Liu, and Han Gao. Is overlain display a right choice for ar navigation? a qualitative study of head-mounted augmented reality surgical navigation on accuracy for large-scale clinical deployment. *CNS Neuroscience & Therapeutics*, 31(1):e70217, 2025.

[50] Christopher M Andrews, Alexander B Henry, Ignacio M Soriano, Michael K Southworth, and Jonathan R Silva. Registration techniques for clinical applications of three-dimensional augmented reality devices. *IEEE journal of translational engineering in health and medicine*, 9:1–14, 2020.

[51] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[52] P.J Edwards, M. Chand, M. Birlo, and D. Stoyanov. The challenge of augmented reality in surgery. In S. Atallah, editor, *Digital Surgery*, chapter 10, pages 121–135. Springer, Cham, 2021.

[53] Manuel Birlo, P.J. Eddie Edwards, Matthew Clarkson, and Danail Stoyanov. Utility of optical see-through head mounted displays in augmented reality-assisted surgery: A systematic review. *Medical Image Analysis*, 77:102361, 2022.

[54] Manuel Birlo, Philip J. Eddie Edwards, Soojeong Yoo, Brian Dromey, Francisco Vasconcelos, Matthew J. Clarkson, and Danail Stoyanov. Cal-tutor: A hololens 2 application for training in obstetric sonography and user motion data recording. *Journal of Imaging*, 9(1), 2023.

[55] Manuel Birlo, Razvan Caramalau, Philip J. "Eddie" Edwards, Brian Dromey, Matthew J. Clarkson, and Danail Stoyanov. HUP-3D: A 3D multi-view synthetic dataset for assisted-egocentric hand-ultrasound-probe pose estimation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15001. Springer Nature Switzerland, October 2024.

[56] Yaoyu Fu, Steven D Schwaitzberg, and Lora Cavuoto. Effects of optical see-through head-mounted display use for simulated laparoscopic surgery. *International Journal of Human–Computer Interaction*, 40(17):4709–4724, 2024.

[57] Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. Apple vision pro: the future of surgery with advances in virtual and augmented reality. *Irish Journal of Medical Science (1971-)*, 193(1):345–346, 2024.

[58] Joshua Olexa, Kevin T Kim, Jordan R Saadon, Maureen Rakovec, Madison Evans, Jonathan Cohen, Jacob Cherian, and Maureen Rokovec. Apple vi-

sion pro augmented reality-assisted minimally invasive surgical treatment of spinal dural arteriovenous fistula. *Cureus*, 16(7), 2024.

[59] M Orione, G Rubegni, R Tartaro, A Alberghina, M Fallico, C Orione, A Russo, GM Tosi, and T Avitabile. Utilization of apple vision pro in ophthalmic surgery: A pilot study. *European Journal of Ophthalmology*, page 11206721241273574, 2024.

[60] Tomoyoshi Okamoto, Shinji Onda, Katsuhiko Yanaga, Naoki Suzuki, and Asaki Hattori. Clinical application of navigation surgery using augmented reality in the abdominal field. *Surgery Today*, 45(4):397–406, Apr 2015.

[61] Fabrizio Cutolo, Umberto Fontana, and Vincenzo Ferrari. Perspective preserving solution for quasi-orthoscopic video see-through hmds. *Technologies*, 6, 01 2018.

[62] Na Guo, Tianmiao Wang, Biao Yang, Lei Hu, Hongsheng Liu, and Yuhan Wang. An online calibration method for microsoft hololens. *IEEE Access*, 7:101795–101803, 2019.

[63] Jannick P Rolland, Richard L Holloway, and Henry Fuchs. Comparison of optical and video see-through, head-mounted displays. In *Telemanipulator and Telepresence Technologies*, volume 2351, pages 293–307. International Society for Optics and Photonics, 1995.

[64] Omid Moshtaghi, Kanwar S Kelley, William B Armstrong, Yaser Ghavami, Jeffery Gu, and Hamid R Djalilian. Using google glass to solve communication and surgical education challenges in the operating room. *The Laryngoscope*, 125(10):2295–2297, 2015.

[65] Rocco Galati, Michele Simone, Graziana Barile, Raffaele De Luca, Carmine Cartanese, and G Grassi. Experimental setup employed in the operating room based on virtual and mixed reality: analysis of pros and cons in open abdomen surgery. *Journal of Healthcare Engineering*, 2020, 2020.

[66] Igor M Sauer, Moritz Queisner, Peter Tang, Simon Moosburner, Ole Hoepfner, Rosa Horner, Rudiger Lohmann, and Johann Pratschke. Mixed reality in visceral surgery: development of a suitable workflow and evaluation of intraoperative use-cases. *Annals of surgery*, 266(5):706–712, 2017.

[67] Marina Carbone, Roberta Piazza, and Sara Condino. Commercially available head-mounted displays are unsuitable for augmented reality surgical guidance: a call for focused research for surgical applications, 2020.

[68] Benish Fida, Fabrizio Cutolo, Gregorio di Franco, Mauro Ferrari, and Vincenzo Ferrari. Augmented reality in open surgery. *Updates in Surgery*, 70(3):389–400, Sep 2018.

[69] James WR Dilley, Archie Hughes-Hallett, Philip J Pratt, Philip H Pucher, Mafalda Camara, Ara W Darzi, and Erik K Mayer. Perfect registration leads to imperfect performance: A randomized trial of multimodal intraoperative image guidance. *Annals of surgery*, 269(2):236–242, 2019.

[70] Alina Solovjova, Benjamin Hatscher, and Christian Hansen. Influence of augmented reality interaction on a primary task for the medical domain. *Mensch und Computer 2019-Workshopband*, 2019.

[71] Darko Katić, Patrick Spengler, Sebastian Bodenstedt, Gregor Castrillon-Oberndorfer, Robin Seeberger, Juergen Hoffmann, Ruediger Dillmann, and Stefanie Speidel. A system for context-aware intraoperative augmented reality in dental implant surgery. *International journal of computer assisted radiology and surgery*, 10(1):101–108, 2015.

[72] Carole Cometti, Christos Païzis, Audrey Casteleira, Guillaume Pons, and Nicolas Babault. Effects of mixed reality head-mounted glasses during 90 minutes of mental and manual tasks on cognitive and physiological functions. *PeerJ*, 6:e5847, 2018.

[73] Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Philip J Devereaux, Jos

Kleijnen, and David Moher. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of internal medicine*, 151(4):W–65, 2009.

[74] Arindam Dey, Mark Billinghurst, Robert W Lindeman, and J Swan. A systematic review of 10 years of augmented reality usability studies: 2005 to 2014. *Frontiers in Robotics and AI*, 5:37, 2018.

[75] L. Chen, T. W. Day, W. Tang, and N. W. John. Recent developments and future challenges in medical mixed reality. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 123–135, 2017.

[76] Martin Eckert, Julia S Volmerg, and Christoph M Friedrich. Augmented reality in medicine: systematic and bibliographic review. *JMIR mHealth and uHealth*, 7(4):e10967, 2019.

[77] Marta Kersten-Oertel, Pierre Jannin, and D Louis Collins. The state of the art of visualization in mixed reality image guided surgery. *Computerized Medical Imaging and Graphics*, 37(2):98–112, 2013.

[78] L. Qian, J. Y. Wu, S. P. DiMaio, N. Navab, and P. Kazanzides. A review of augmented reality in robotic-assisted surgery. *IEEE Transactions on Medical Robotics and Bionics*, 2(1):1–16, 2020.

[79] Sylvain Bernhardt, Stéphane A Nicolau, Luc Soler, and Christophe Doignon. The status of augmented reality in laparoscopic surgery as of 2016. *Medical image analysis*, 37:66–90, 2017.

[80] Carl Laverdière, Jason Corban, Jason Khoury, Susan Mengxiao Ge, Justin Schupbach, Edward J Harvey, Rudy Reindl, and Paul A Martineau. Augmented reality in orthopaedics: a systematic review and a window on future possibilities. *The Bone & Joint Journal*, 101(12):1479–1488, 2019.

[81] Lukas Jud, Javad Fotouhi, Octavian Andronic, Alexander Aichmair, Greg Osgood, Nassir Navab, and Mazda Farshad. Applicability of augmented reality in orthopedic surgery–a systematic review. *BMC musculoskeletal disorders*, 21(1):1–13, 2020.

[82] Jens T Verhey, Jack M Haglin, Erik M Verhey, and David E Hartigan. Virtual, augmented, and mixed reality applications in orthopedic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 16(2):e2067, 2020.

[83] Sandra Barteit, Lucia Lanfermann, Till Bärnighausen, Florian Neuhann, Claudia Beiersmann, et al. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. *JMIR serious games*, 9(3):e29080, 2021.

[84] Xuanhui Xu, Eleni Mangina, and Abraham G Campbell. Hmd-based virtual and augmented reality in medical education: a systematic review. *Frontiers in Virtual Reality*, 2:692103, 2021.

[85] Kyle McCloskey, Ryan Turlip, Hasan S. Ahmad, Yohannes G. Ghenbot, Daksh Chauhan, and Jang W. Yoon. Virtual and augmented reality in spine surgery: A systematic review. *World Neurosurgery*, 173:96–107, 2023.

[86] Mingxiao Tu, Hoijoon Jung, Jinman Kim, and Andre Kyme. Head-mounted displays in context-aware systems for open surgery: A state-of-the-art review. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[87] David G Armstrong, Timothy M Rankin, Nicholas A Giovinco, Joseph L Mills, and Yoky Matsuoka. A heads-up display for diabetic limb salvage surgery: a view through the google looking glass. *Journal of diabetes science and technology*, 8(5):951–956, 2014.

[88] Brent A Ponce, Mariano E Menendez, Lasun O Oladeji, Charles T Fryberger, and Phani K Dantuluri. Emerging technology in surgical education: combin-

ing real-time augmented reality and wearable computing devices. *Orthopedics*, 37(11):751–757, 2014.

[89] Xiaojun Chen, Lu Xu, Yiping Wang, Huixiang Wang, Fang Wang, Xiangsen Zeng, Qiugen Wang, and Jan Egger. Development of a surgical navigation system based on augmented reality using an optical see-through head-mounted display. *Journal of biomedical informatics*, 55:124–131, 2015.

[90] Huixiang Wang, Fang Wang, Anthony Peng Yew Leong, Lu Xu, Xiaojun Chen, and Qiugen Wang. Precision insertion of percutaneous sacroiliac screws using a novel augmented reality-based navigation system: a pilot study. *International orthopaedics*, 40(9):1941–1947, 2016.

[91] James Stewart and Mark Billinghurst. A wearable navigation display can improve attentiveness to the surgical field. *International journal of computer assisted radiology and surgery*, 11(6):1193–1200, 2016.

[92] Takafumi Hiranaka, Takaaki Fujishiro, Yuichi Hida, Yosaku Shibata, Masanori Tsubosaka, Yuta Nakanishi, Kenjiro Okimura, and Harunobu Uemoto. Augmented reality: the use of the picolinker smart glasses improves wire insertion under fluoroscopy. *World journal of orthopedics*, 8(12):891, 2017.

[93] Shahram Jalaliniya, Thomas Pederson, and Diako Mardanbegi. A wearable personal assistant for surgeons: Design, evaluation, and future prospects. *EAI Endorsed Transactions on Pervasive Health and Technology*, 3(12), 2017.

[94] Mathias Unberath, Javad Fotouhi, Jonas Hajek, Andreas Maier, Greg Osgood, Russell Taylor, Mehran Armand, and Nassir Navab. Augmented reality-based feedback for technician-in-the-loop c-arm repositioning. *Healthcare technology letters*, 5(5):143–147, 2018.

[95] Houssam El-Hariri, Prashant Pandey, Antony J Hodgson, and Rafeef Garbi. Augmented reality visualisation for orthopaedic surgical guidance with pre-

and intra-operative multimodal image data fusion. *Healthcare Technology Letters*, 5(5):189–193, 2018.

[96] Gerard Deib, Alex Johnson, Mathias Unberath, Kevin Yu, Sebastian Andress, Long Qian, Gregory Osgood, Nassir Navab, Ferdinand Hui, and Philippe Gailloud. Image guided percutaneous spine procedures using an optical see-through head mounted display: proof of concept and rationale. *Journal of neurointerventional surgery*, 10(12):1187–1191, 2018.

[97] Sebastian Andress, Alex Johnson M.D., Mathias Unberath, Alexander F. Winkler, Kevin Yu, Javad Fotouhi, Simon Weidert M.D., Greg M. Osgood M.D., and Nassir Navab. On-the-fly augmented reality for orthopedic surgery using a multimodal fiducial. *Journal of Medical Imaging*, 5(2):1 – 12, 2018.

[98] Sara Condino, Giuseppe Turini, Paolo D Parchi, Rosanna M Viglialoro, Nicola Piolanti, Marco Gesi, Mauro Ferrari, and Vincenzo Ferrari. How to build a patient-specific hybrid simulator for orthopaedic open surgery: benefits and limits of mixed-reality using the microsoft hololens. *Journal of Healthcare Engineering*, 2018, 2018.

[99] Jacob T Gibby, Samuel A Swenson, Steve Cvetko, Raj Rao, and Ramin Javan. Head-mounted display augmented reality to guide pedicle screw placement utilizing computed tomography. *International journal of computer assisted radiology and surgery*, 14(3):525–535, 2019.

[100] Marcelo E de Oliveira, Henrique G Debarba, Alexandre Lädermann, Sylvain Chagué, and Caecilia Charbonnier. A hand-eye calibration method for augmented reality applied to computer-assisted orthopedic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 15(2):e1969, 2019.

[101] Jacob Aaskov, Gregory N Kawchuk, Kenton D Hamaluik, Pierre Boulanger, and Jan Hartvigsen. X-ray vision: the accuracy and repeatability of a tech-

nology that allows clinicians to see spinal x-rays superimposed on a person's back. *PeerJ*, 7:e6333, 2019.

[102] Florentin Liebmann, Simon Roner, Marco von Atzigen, Davide Scaramuzza, Reto Sutter, Jess Snedeker, Mazda Farshad, and Philipp Fürnstahl. Pedicle screw navigation using surface digitization on the microsoft hololens. *International journal of computer assisted radiology and surgery*, 14(7):1157–1165, 2019.

[103] Javad Fotouhi, Mathias Unberath, Tianyu Song, Wenhao Gu, Alex Johnson, Greg Osgood, Mehran Armand, and Nassir Navab. Interactive flying frustums (iffs): spatially aware surgical data visualization. *International journal of computer assisted radiology and surgery*, 14(6):913–922, 2019.

[104] Javad Fotouhi, Mathias Unberath, Tianyu Song, Jonas Hajek, Sing Chun Lee, Bastian Bier, Andreas Maier, Greg Osgood, Mehran Armand, and Nassir Navab. Co-localized augmented human and x-ray observers in collaborative surgical ecosystem. *International journal of computer assisted radiology and surgery*, 14(9):1553–1563, 2019.

[105] Piotr Pietruski, Marcin Majak, Ewelina Świątek-Najwer, Magdalena Żuk, Michał Popek, Janusz Jaworowski, and Maciej Mazurek. Supporting fibula free flap harvest with augmented reality: A proof-of-concept study. *The Laryngoscope*, 130(5):1173–1179, 2020.

[106] B Laguna, K Livingston, R Brar, J Jagodzinski, N Pandya, C Sabatini, and J Courtier. Assessing the value of a novel augmented reality application for presurgical planning in adolescent elbow fractures. front. *Virtual Real. 1: 528810. doi: 10.3389/frvir*, 2020.

[107] Jacob Gibby, Steve Cvetko, Ramin Javan, Ryan Parr, and Wendell Gibby. Use of augmented reality for image-guided spine procedures. *European Spine Journal*, 29(8):1823–1832, 2020.

[108] Wenhao Gu, Kinjal Shah, Jonathan Knopf, Nassir Navab, and Mathias Unberath. Feasibility of image-based augmented reality guidance of total shoulder arthroplasty using microsoft hololens 1. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–10, 2020.

[109] Arnd Fredrik Viehöfer, Stephan Hermann Wirth, Stefan Michael Zimmermann, Laurenz Jaberg, Cyrill Dennler, Philipp Fürnstahl, and Mazda Farshad. Augmented reality guided osteotomy in hallux valgus correction. *BMC Musculoskeletal Disorders*, 21(1):1–6, 2020.

[110] Cyrill Dennler, Laurenz Jaberg, José Spirig, Christoph Agten, Tobias Götschi, Philipp Fürnstahl, and Mazda Farshad. Augmented reality-based navigation increases precision of pedicle screw insertion. *Journal of orthopaedic surgery and research*, 15:1–8, 2020.

[111] Philipp Kriechling, Simon Roner, Florentin Liebmann, Fabio Casari, Philipp Fürnstahl, and Karl Wieser. Augmented reality for base plate component placement in reverse total shoulder arthroplasty: a feasibility study. *Archives of orthopaedic and trauma surgery*, pages 1–7, 2020.

[112] Keitaro Matsukawa and Yoshiyuki Yato. Smart glasses display device for fluoroscopically guided minimally invasive spinal instrumentation surgery: a preliminary study. *Journal of Neurosurgery: Spine*, 1(aop):1–6, 2020.

[113] Michael A Lin, Alexa F Siu, Jung Hwa Bae, Mark R Cutkosky, and Bruce L Daniel. Holoneedle: augmented reality guidance system for needle placement investigating the advantages of three-dimensional needle shape reconstruction. *IEEE Robotics and Automation Letters*, 3(4):4156–4162, 2018.

[114] Ming-Long Wu, Jong-Chih Chien, Chieh-Tsai Wu, and Jiann-Der Lee. An augmented reality system using improved-iterative closest point algorithm for on-patient medical image visualization. *Sensors*, 18(8):2505, 2018.

[115] Faraz Mahmood, Eitezaz Mahmood, Robert Gregory Dorfman, John Mitchell, Feroze-Udin Mahmood, Stephanie B Jones, and Robina Matyal. Augmented reality and ultrasound education: initial experience. *Journal of cardiothoracic and vascular anesthesia*, 32(3):1363–1367, 2018.

[116] Edgar Rojas-Muñoz, Maria Eugenia Cabrera, Daniel Andersen, Voicu Popescu, Sherri Marley, Brian Mullis, Ben Zarzaur, and Juan Wachs. Surgical telementoring without encumbrance: a comparative study of see-through augmented reality-based approaches. *Annals of surgery*, 270(2):384–389, 2019.

[117] Ruotong Li, Weixin Si, Xiangyun Liao, Qiong Wang, Reinhard Klein, and Pheng-Ann Heng. Mixed reality based respiratory liver tumor puncture navigation. *Computational Visual Media*, 5(4):363–374, 2019.

[118] Egidijus Pelanis, Rahul P Kumar, Davit L Aghayan, Rafael Palomar, Åsmund A Fretland, Henrik Brun, Ole Jakob Elle, and Bjørn Edwin. Use of mixed reality for improved spatial understanding of liver anatomy. *Minimally Invasive Therapy & Allied Technologies*, 29(3):154–160, 2020.

[119] Zeyang Zhou, Zhiyong Yang, Shan Jiang, Fujun Zhang, Huzheng Yan, and Xiaodong Ma. Surgical navigation system for low-dose-rate brachytherapy based on mixed reality. *IEEE Computer Graphics and Applications*, 2020.

[120] Hasaneen Fathy Al Janabi, Abdullatif Aydin, Sharanya Palaneer, Nicola Macchione, Ahmed Al-Jabir, Muhammad Shamim Khan, Prokar Dasgupta, and Kamran Ahmed. Effectiveness of the hololens mixed-reality headset in minimally invasive surgery: a simulation-based feasibility study. *Surgical Endoscopy*, 34(3):1143–1149, 2020.

[121] Guan Li, Jie Dong, Jinbao Wang, Dongbing Cao, Xin Zhang, Zhiqiang Cao, and Guangming Lu. The clinical application value of mixed-reality-assisted surgical navigation for laparoscopic nephrectomy. *Cancer Medicine*, 9(15):5480–5489, 2020.

[122] Jang W Yoon, Robert E Chen, Phillip K Han, Phong Si, William D Freeman, and Stephen M Pirris. Technical feasibility and safety of an intraoperative head-up display device during spine instrumentation. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 13(3):e1770, 2017.

[123] Christof Karmonik, Saba N Elias, Jonathan Y Zhang, Orlando Diaz, Richard P Klucznik, Robert G Grossman, and Gavin W Britz. Augmented reality with virtual cerebral aneurysms: A feasibility study. *World neurosurgery*, 119:e617–e622, 2018.

[124] Taylor Frantz, Bart Jansen, Johnny Duerinck, and Jef Vandemeulebroucke. Augmenting microsoft's hololens with vuforia tracking for neuronavigation. *Healthcare technology letters*, 5(5):221–225, 2018.

[125] Nhu Q Nguyen, Jillian Cardinell, Joel M Ramjist, Philips Lai, Yuta Dobashi, Daipayan Guha, Dimitrios Androutsos, and Victor XD Yang. An augmented reality system characterization of placement accuracy in neurosurgery. *Journal of Clinical Neuroscience*, 72:392–396, 2020.

[126] Zhen-yu Zhang, Wen-chao Duan, Ruo-kun Chen, Feng-jiang Zhang, Bin Yu, Yun-bo Zhan, Ke Li, Hai-biao Zhao, Tao Sun, Yu-chen Ji, et al. Preliminary application of mxed reality in neurosurgery: Development and evaluation of a new intraoperative procedure. *Journal of Clinical Neuroscience*, 67:234–238, 2019.

[127] Florian Heinrich, Luisa Schwenderling, Mathias Becker, Martin Skalej, and Christian Hansen. Holoinjection: augmented reality support for ct-guided spinal needle injections. *Healthcare Technology Letters*, 6(6):165–171, 2019.

[128] ZM Baum, Andras Lasso, Sarah Ryan, Tamas Ungi, Emily Rae, Boris Zevin, Ron Levy, and Gabor Fichtinger. Augmented reality training platform for neurosurgical burr hole localization. *J Med Robot Res*, pages 194–2001, 2020.

[129] Jason I Liounakos, Timur Urakov, and Michael Y Wang. Head-up display assisted endoscopic lumbar discectomy—a technical note. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 16(3):e2089, 2020.

[130] Xuetong Sun, Sarah B Murthi, Gary Schwartzbauer, and Amitabh Varshney. High-precision 5 DoF tracking and visualization of catheter placement in EVD of the brain using AR. *ACM Transactions on Computing for Healthcare*, 1(2):1–18, 2020.

[131] Naoki Kaneko, Makoto Sato, Taro Takeshima, Yoshihide Sehara, and Eiju Watanabe. Ultrasound-guided central venous catheterization using an optical see-through head-mounted display: A pilot study. *Journal of Clinical Ultrasound*, 44(8):487–491, 2016.

[132] Ivo Kuhlemann, Markus Kleemann, Philipp Jauer, Achim Schweikard, and Floris Ernst. Towards x-ray free endovascular interventions–using hololens for on-line holographic visualisation. *Healthcare technology letters*, 4(5):184–187, 2017.

[133] Zeyang Zhou, Zhiyong Yang, Shan Jiang, Fujun Zhang, and Huzheng Yan. Design and validation of a surgical navigation system for brachytherapy based on mixed reality. *Medical physics*, 46(8):3709–3718, 2019.

[134] Paweł Rynio, Jan Witowski, Jakub Kamiński, Jakub Serafin, Arkadiusz Kazimierczak, and Piotr Gutowski. Holographically-guided endovascular aneurysm repair. *Journal of Endovascular Therapy*, 26(4):544–547, 2019.

[135] Edgar Rojas-Muñoz, Maria E Cabrera, Chengyuan Lin, Daniel Andersen, Voicu Popescu, Kathryn Anderson, Ben L Zarzaur, Brian Mullis, and Juan P Wachs. The system for telementoring with augmented reality (star): A head-mounted display to improve surgical coaching and confidence in remote areas. *Surgery*, 2020.

[136] Brian J Park, Nicholas R Perkons, Enri Profka, Omar Johnson, Christopher Morley, Scott Appel, Gregory J Nadolski, Stephen J Hunt, and Terence P Gade. Three-dimensional augmented reality visualization informs locoregional therapy in a translational model of hepatocellular carcinoma. *Journal of Vascular and Interventional Radiology*, 31(10):1612–1618, 2020.

[137] Helena Catarina Margarido Mendes, Cátia Isabel Andrade Botelho Costa, Nuno André da Silva, Francisca Pais Leite, Augusto Esteves, and Daniel Simões Lopes. Piñata: Pinpoint insertion of intravenous needles via augmented reality training assistance. *Computerized Medical Imaging and Graphics*, 82:101731, 2020.

[138] Jene W Meulstee, Johan Nijsink, Ruud Schreurs, Luc M Verhamme, Tong Xi, Hans HK Delye, Wilfred A Borstlap, and Thomas JJ Maal. Toward holographic-guided surgery. *Surgical innovation*, 26(1):86–94, 2019.

[139] Jong-Chih Chien, Yao-Ren Tsai, Chieh-Tsai Wu, and Jiann-Der Lee. Hololens-based ar system with a robust point set registration algorithm. *Sensors*, 19(16):3555, 2019.

[140] Thomas Boillat, Peter Grantcharov, and Homero Rivas. Increasing completion rate and benefits of checklists: Prospective evaluation of surgical safety checklists with smart glasses. *JMIR mHealth and uHealth*, 7(4):e13447, 2019.

[141] David Dallas-Orr, Yordan Penev, Robert Schultz, and Jesse Courtier. Comparing computed tomography–derived augmented reality holograms to a standard picture archiving and communication systems viewer for presurgical planning: Feasibility study. *JMIR Perioperative Medicine*, 3(2):e18367, 2020.

[142] Javier A Luzon, Bojan V Stimec, Arne O Bakka, Bjørn Edwin, and Dejan Ignjatovic. Value of the surgeon's sightline on hologram registration and

targeting in mixed reality. *International Journal of Computer Assisted Radiology and Surgery*, 15(12):2027–2039, 2020.

[143] João Cartucho, David Shapira, Hutan Ashrafian, and Stamatia Giannarou. Multimodal mixed reality visualisation for intraoperative surgical guidance. *International journal of computer assisted radiology and surgery*, 15(5):819–826, 2020.

[144] Rahul Prasanna Kumar, Egidijus Pelanis, Robin Bugge, Henrik Brun, Rafael Palomar, Davit L Aghayan, smund Avdem Fretland, Bjørn Edwin, and Ole Jakob Elle. Use of mixed reality for surgery planning: Assessment and development workflow. *Journal of Biomedical Informatics: X*, 8:100077, 2020.

[145] Cara A Liebert, Mohamed A Zayed, Oliver Aalami, Jennifer Tran, and James N Lau. Novel use of google glass for procedural wireless vital sign monitoring. *Surgical innovation*, 23(4):366–373, 2016.

[146] Tianyu Song, Chenglin Yang, Omid Dianat, and Ehsan Azimi. Endodontic guided treatment using augmented reality on a head-mounted display system. *Healthcare Technology Letters*, 5(5):201–207, 2018.

[147] Yaxuan Zhou, Paul Yoo, Yingru Feng, Aditya Sankar, Alireza Sadr, and Eric J Seibel. Towards ar-assisted visualisation and guidance for imaging of dental decay. *Healthcare technology letters*, 6(6):243–248, 2019.

[148] Sobia Zafar and Jessica Joanna Zachar. Evaluation of holohuman augmented reality application as a novel educational tool in dentistry. *European Journal of Dental Education*, 24(2):259–265, 2020.

[149] Qiming Li, Chen Huang, Shengqing Lv, Zeyu Li, Yimin Chen, and Lizhuang Ma. An human-computer interactive augmented reality system for coronary artery diagnosis planning and training. *Journal of medical systems*, 41(10):159, 2017.

[150] Yi-bo Zou, Yi-min Chen, Ming-ke Gao, Quan Liu, Si-yu Jiang, Jia-hui Lu, Chen Huang, Ze-yu Li, and Dian-hua Zhang. Coronary heart disease preoperative gesture interactive diagnostic system based on augmented reality. *Journal of medical systems*, 41(8):126, 2017.

[151] H Brun, RAB Bugge, LKR Suther, S Birkeland, R Kumar, E Pelanis, and OJ Elle. Mixed reality holograms for heart surgery planning: first user experience in congenital heart disease. *European Heart Journal-Cardiovascular Imaging*, 20(8):883–888, 2019.

[152] Jun Liu, Subhi J Al'Aref, Gurpreet Singh, Alexandre Caprio, Amir Ali Amiri Moghadam, Sun-Joo Jang, S Chiu Wong, James K Min, Simon Dunham, and Bobak Mosadegh. An augmented reality system for image guidance of transcatheter procedures for structural heart disease. *PloS one*, 14(7), 2019.

[153] Edgar Rojas-Muñoz, Chengyuan Lin, Natalia Sanchez-Tamayo, Maria Eugenia Cabrera, Daniel Andersen, Voicu Popescu, Juan Antonio Barragan, Ben Zarzaur, Patrick Murphy, Kathryn Anderson, et al. Evaluation of an augmented reality platform for austere surgical telementoring: a randomized controlled crossover study in cricothyroidotomies. *NPJ Digital Medicine*, 3(1):1–9, 2020.

[154] Joshua J Gnanasegaram, Regina Leung, and Jason A Beyea. Evaluating the effectiveness of learning ear anatomy using holographic models. *Journal of Otolaryngology-Head & Neck Surgery*, 49(1):1–8, 2020.

[155] Claudia Scherl, Johanna Stratemeier, Celine Karle, Nicole Rotter, Jürgen Hesser, Lena Huber, Andre Dias, Oliver Hoffmann, Philipp Riffel, Stefan O Schoenberg, et al. Augmented reality with hololens in parotid surgery: how to assess and to improve accuracy. *European Archives of Oto-Rhino-Laryngology*, pages 1–11, 2020.

[156] Francis X Creighton, Mathias Unberath, Tianyu Song, Zhuokai Zhao, Mehran Armand, and John Carey. Early feasibility studies of augmented

reality navigation for lateral skull base surgery. *Otology & Neurotology*, 41(7):883–888, 2020.

[157] Daisuke Mitsuno, Koichi Ueda, Tomoki Itamiya, Takashi Nuri, and Yuki Otsuki. Intraoperative evaluation of body surface improvement by an augmented reality system that a clinician can modify. *Plastic and Reconstructive Surgery Global Open*, 5(8), 2017.

[158] Philip Pratt, Matthew Ives, Graham Lawton, Jonathan Simmons, Nasko Radev, Liana Spyropoulou, and Dimitri Amiras. Through the hololens™ looking glass: augmented reality for extremity reconstruction surgery using 3d vascular models with perforating vessels. *European radiology experimental*, 2(1):2, 2018.

[159] Taoran Jiang, Dewang Yu, Yuqi Wang, Tao Zan, Shuyi Wang, and Qingfeng Li. Hololens-based vascular localization system: Precision evaluation study with a three-dimensional printed model. *Journal of medical Internet research*, 22(4):e16852, 2020.

[160] Piotr Pietruski, Marcin Majak, Ewelina Świątek-Najwer, Magdalena Żuk, Michał Popek, Maciej Mazurek, Marta Świecka, and Janusz Jaworowski. Supporting mandibular resection with intraoperative navigation utilizing augmented reality technology–a proof of concept study. *Journal of Cranio-Maxillofacial Surgery*, 47(6):854–859, 2019.

[161] Antonio Pepe, Gianpaolo Francesco Trotta, Peter Mohr-Ziak, Christina Gsaxner, Jürgen Wallner, Vitoantonio Bevilacqua, and Jan Egger. A markerless registration approach for mixed reality–aided maxillofacial surgery: a pilot evaluation. *Journal of digital imaging*, 32(6):1008–1018, 2019.

[162] Qichang Sun, Yongfeng Mai, Rong Yang, Tong Ji, Xiaoyi Jiang, and Xiaojun Chen. Fast and accurate online calibration of optical see-through head-mounted display for ar-based surgical navigation using microsoft

hololens. *International Journal of Computer Assisted Radiology and Surgery*, 15(11):1907–1919, 2020.

[163] H Borgmann, M Rodríguez Socarrás, J Salem, I Tsaur, J Gomez Rivas, E Barret, and L Tortolero. Feasibility and safety of augmented reality-assisted urological surgery using smartglass. *World Journal of Urology*, 6(35):967–972, 2016.

[164] Ryan M Dickey, Neel Srikishen, Larry I Lipshultz, Philippe E Spiess, Rafael E Carrion, and Tariq S Hakky. Augmented reality assisted surgery: a urologic training tool. *Asian journal of andrology*, 18(5):732, 2016.

[165] DS Schoeb, J Schwarz, S Hein, D Schlager, PF Pohlmann, A Frankenschmidt, C Gratzke, and A Miernik. Mixed reality for teaching catheter placement to medical students: a randomized single-blinded, prospective trial. *BMC medical education*, 20(1):1–8, 2020.

[166] Long Qian, Anton Deguet, and Peter Kazanzides. Arssist: augmented reality on a head-mounted display for the first assistant in robotic surgery. *Healthcare technology letters*, 5(5):194–200, 2018.

[167] Javad Fotouhi, Tianyu Song, Arian Mehrfard, Giacomo Taylor, Qiaochu Wang, Fengfan Xian, Alejandro Martin-Gomez, Bernhard Fuerst, Mehran Armand, Mathias Unberath, et al. Reflective-ar display: An interaction methodology for virtual-to-real alignment in medical robotics. *IEEE Robotics and Automation Letters*, 5(2):2722–2729, 2020.

[168] Lianne M Wellens, Jene Meulstee, Cornelis P van de Ven, CEJ Terwisscha van Scheltinga, Annemieke S Littooij, Marry M van den Heuvel-Eibrink, Marta Fiocco, Anne C Rios, Thomas Maal, and Marc HWA Wijnen. Comparison of 3-dimensional and augmented reality kidney models with conventional imaging data in the preoperative assessment of children with wilms tumors. *JAMA network open*, 2(4):e192633–e192633, 2019.

[169] Matthijs Fitski, Jene W Meulstee, Annemieke S Littooij, Cornelis P van de Ven, Alida FW van der Steeg, and Marc HWA Wijnen. Mri-based 3-dimensional visualization workflow for the preoperative planning of nephron-sparing surgery in wilms' tumor surgery: A pilot study. *Journal of Healthcare Engineering*, 2020, 2020.

[170] Ming Li, Reza Seifabadi, Dilara Long, Quirina De Ruiter, Nicole Varble, Rachel Hecht, Ayele H Negussie, Venkatesh Krishnasamy, Sheng Xu, and Bradford J Wood. Smartphone-versus smartglasses-based augmented reality (ar) for percutaneous needle interventions: system accuracy and feasibility study. *International Journal of Computer Assisted Radiology and Surgery*, 15(11):1921–1930, 2020.

[171] Ezequiel Roberto Zorzal, José Miguel Campos Gomes, Maurício Sousa, Pedro Belchior, Pedro Garcia da Silva, Nuno Figueiredo, Daniel Simões Lopes, and Joaquim Jorge. Laparoscopy with augmented reality adaptations. *Journal of biomedical informatics*, 107:103463, 2020.

[172] Paul D Schlosser, Tobias Grundgeiger, Penelope M Sanderson, and Oliver Happel. An exploratory clinical evaluation of a head-worn display based multiple-patient monitoring application: impact on supervising anesthesiologists' situation awareness. *Journal of clinical monitoring and computing*, 33(6):1119–1127, 2019.

[173] Jan Egger, Christina Gsaxner, Gijs Luijten, Jianxu Chen, Xiaojun Chen, Jiang Bian, Jens Kleesiek, Behrus Puladi, et al. Is the apple vision pro the ultimate display? a first perspective and survey on entering the wonderland of precision medicine. *JMIR Serious Games*, 12(1):e52785, 2024.

[174] Kevin Cleary and Terry M Peters. Image-guided interventions: technology review and clinical applications. *Annual review of biomedical engineering*, 12:119–142, 2010.

[175] Shahram Jalaliniya and Thomas Pederson. Designing wearable personal assistants for surgeons: An egocentric approach. *IEEE Pervasive Computing*, 14(3):22–31, 2015.

[176] Yinlong Liu, Zhijian Song, and Manning Wang. A new robust markerless method for automatic image-to-patient registration in image-guided neurosurgery system. *Computer Assisted Surgery*, 22(sup1):319–325, 2017.

[177] Medha V Wyawahare, Pradeep M Patil, Hemant K Abhyankar, et al. Image registration techniques: an overview. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2(3):11–28, 2009.

[178] A Seginer. Rigid-body point-based registration: The distribution of the target registration error when the fiducial registration errors are given. *Medical image analysis*, 15(4):397–413, 2011.

[179] Bill Papantoniou, M Soegaard, JR Lupton, M Goktürk, and D Trepess. The glossary of human computer interaction. *Online source:* `https://www.interaction-design.org/literature/book/the-glossary-of-human-computer-interaction` *[2019-04-23]*, 2016.

[180] Bethany R Lowndes and M Susan Hallbeck. Overview of human factors and ergonomics in the or, with an emphasis on minimally invasive surgeries. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 24(3):308–317, 2014.

[181] M. A. Livingston. Evaluating human factors in augmented reality systems. *IEEE Computer Graphics and Applications*, 25(6):6–9, 2005.

[182] Weidong Huang, Leila Alem, and Mark A Livingston. *Human factors in augmented reality environments*. Springer Science & Business Media, 2012.

[183] Arthur Tang, Ji Zhou, and Charles Owen. Evaluation of calibration procedures for optical see-through head-mounted displays. In *The Second IEEE*

and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings., pages 161–168. IEEE, 2003.

[184] Center for Devices and Radiological Health. Applying human factors and usability engineering to medical devices. *Guidance for Industry and Food and Drug Administration Staff. FDA-2011-D-0469*, 2016.

[185] Yan Zuo, Taoran Jiang, Jiansheng Dou, Dewang Yu, Zaphlene Nyakuru Ndaro, Yunxiao Du, Qingfeng Li, Shuyi Wang, and Gang Huang. A novel evaluation model for a mixed-reality surgical navigation system: Where microsoft hololens meets the operating room. *Surgical Innovation*, 27(2):193–202, 2020.

[186] Jayfus T Doswell and Anna Skinner. Augmenting human cognition with adaptive augmented reality. In *International Conference on Augmented Cognition*, pages 104–113. Springer, 2014.

[187] Marta Kersten-Oertel, Pierre Jannin, and D Louis Collins. Dvv: A taxonomy for mixed reality visualization in image guided surgery. *IEEE Transactions on Visualization and Computer Graphics*, 2(18):332–352, 2012.

[188] Ehsan Azimi, Camilo Molina, Alexander Chang, Judy Huang, Chien-Ming Huang, and Peter Kazanzides. Interactive training and operation ecosystem for surgical tasks in mixed reality. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 20–29. Springer, 2018.

[189] WX Chen, XY Cui, J Zheng, JM Zhang, S Chen, and YD Yao. Gaze gestures and their applications in human-computer interaction with a head-mounted display. *arXiv preprint arXiv:1910.07428*, 2019.

[190] Thibault Louis, Jocelyne Troccaz, Amélie Rochet-Capellan, Nady Hoyek, and François Bérard. When high fidelity matters: Ar and vr improve the learning of a 3d object. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–9, 2020.

[191] Agnes Wittek, Brigitte Strizek, and Florian Recker. Innovations in ultrasound training in obstetrics. *Archives of Gynecology and Obstetrics*, pages 1–10, 2024.

[192] Kwok-Yin Leung. Applications of advanced ultrasound technology in obstetrics. *Diagnostics*, 11(7):1217, 2021.

[193] Sophia Bano, Brian Dromey, Francisco Vasconcelos, Raffaele Napolitano, Anna L David, Donald M Peebles, and Danail Stoyanov. Autofb: automating fetal biometry estimation from standard ultrasound planes. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, pages 228–238. Springer, 2021.

[194] L.J. Salomon, Z. Alfirevic, F. Da Silva Costa, R.L. Deter, F. Figueras, T. Ghi, P. Glanc, A. Khalil, W. Lee, R. Napolitano, A. Papageorghiou, A. Sotiriadis, J. Stirnemann, A. Toi, and G. Yeo. Isuog practice guidelines: ultrasound assessment of fetal biometry and growth. *Ultrasound in Obstetrics & Gynecology*, 53(6):715–723, 2019.

[195] Juncheng Guo, Guanghua Tan, Fan Wu, Huaxuan Wen, and Kenli Li. Fetal ultrasound standard plane detection with coarse-to-fine multi-task learning. *IEEE Journal of Biomedical and Health Informatics*, 27(10):5023–5031, 2022.

[196] Tobias Todsen, Morten Lind Jensen, Martin Grønnebæk Tolsgaard, Beth Härstedt Olsen, Birthe Merete Henriksen, Jens Georg Hillingsø, Lars Konge, and Charlotte Ringsted. Transfer from point-of-care Ultrasonography training to diagnostic performance on patients—a randomized controlled trial. *The American Journal of Surgery*, 211(1):40–45, 1 2016.

[197] Shu-Chen Liao, Shih-Chieh Shao, Shi-Ying Gao, and Edward Chia-Cheng Lai. Augmented reality visualization for ultrasound-guided interventions: a

pilot randomized crossover trial to assess trainee performance and cognitive load. *BMC Medical Education*, 24(1):1058, 2024.

[198] MICHAEL B KIMMEY, FRED E SILVERSTEIN, RODGER C HAGGITT, WILLIAM P SHUMAN, LAURENCE A MACK, CHARLES A ROHRMANN, ALBERT A MOSS, and DONALD W FRANKLIN. Cross-sectional imaging method a system to compare ultrasound, computed tomography, and magnetic resonance with histologic findings. *Investigative Radiology*, 22(3):227–231, 1987.

[199] Bing Wu, Roberta L Klatzky, and George D Stetten. Mental visualization of objects from cross-sectional images. *Cognition*, 123(1):33–49, 2012.

[200] Bahbibi Rahmatullah, Ippokratis Sarris, Aris Papageorghiou, and J. Alison Noble. Quality control of fetal ultrasound images: Detection of abdomen anatomical landmarks using adaboost. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 6–9, 2011.

[201] Hao Chen, Lingyun Wu, Qi Dou, Jing Qin, Shengli Li, Jie-Zhi Cheng, Dong Ni, and Pheng-Ann Heng. Ultrasound standard plane detection using a composite neural network framework. *IEEE Transactions on Cybernetics*, 47(6):1576–1586, 2017.

[202] I Sarris, C Ioannou, M Dighe, A Mitidieri, M Oberto, W Qingqing, J Shah, S Sohoni, W Al Zidjali, L Hoch, et al. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. *Ultrasound in obstetrics & gynecology*, 38(6):681–687, 2011.

[203] Brian J Park, Stephen J Hunt, Gregory J Nadolski, and Terence P Gade. Augmented reality improves procedural efficiency and reduces radiation dose for ct-guided lesion targeting: a phantom study using hololens 2. *Scientific reports*, 10(1):18620, 2020.

[204] Ming Li, Sherif Mehralivand, Sheng Xu, Nicole Varble, Ivane Bakhutashvili, Sandeep Gurram, Peter A Pinto, Peter L Choyke, Bradford J Wood, and Baris

Turkbey. Hololens augmented reality system for transperineal free-hand prostate procedures. *Journal of Medical Imaging*, 10(2):025001–025001, 2023.

[205] Sean P Shevlin, Lloyd Turbitt, David Burckett-St Laurent, Alan J Macfarlane, Simeon West, and James S Bowness. Augmented reality in ultrasound-guided regional anaesthesia: an exploratory study on models with potential implications for training. *Cureus*, 15(7), 2023.

[206] Christoph Rüger, Markus A Feufel, Simon Moosburner, Christopher Özbek, Johann Pratschke, and Igor M Sauer. Ultrasound in augmented reality: a mixed-methods evaluation of head-mounted displays in image-guided interventions. *International Journal of Computer Assisted Radiology and Surgery*, 15:1895–1905, 2020.

[207] Bing Wu, Roberta L Klatzky, and George Stetten. Visualizing 3d objects from 2d cross sectional images displayed in-situ versus ex-situ. *Journal of Experimental Psychology: Applied*, 16(1):45, 2010.

[208] Nazlee Sharmin, Ava K Chow, and Sharla King. Effect of teaching tools in spatial understanding in health science education: A systematic review. *Canadian Medical Education Journal*, 14(4):70–88, 2023.

[209] Thomas Saliba and Sanjiva Pather. The use of virtual reality and augmented reality in ultrasound education, a narrative review of the literature. *Journal of Clinical Ultrasound*, 2024.

[210] Brian P Dromey, Donald M Peebles, and Danail V Stoyanov. A Systematic Review and Meta-analysis of the Use of High-Fidelity Simulation in Obstetric Ultrasound. *Simulation in healthcare : journal of the Society for Simulation in Healthcare*, 7 2020.

[211] Qianqian Cai, Chang Peng, Jian-Yu Lu, Juan C Prieto, Alan J Rosenbaum, Jeffrey SA Stringer, and Xiaoning Jiang. Performance enhanced ultrasound

probe tracking with a hemispherical marker rigid body. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 68(6):2155–2163, 2021.

[212] Christopher Mela, Francis Papay, and Yang Liu. Novel multimodal, multiscale imaging system with augmented reality. *Diagnostics*, 11(3):441, 2021.

[213] Xinyang Liu, William Plishker, and Raj Shekhar. Hybrid electromagnetic-aruco tracking of laparoscopic ultrasound transducer in laparoscopic video. *Journal of Medical Imaging*, 8(1):015001–015001, 2021.

[214] Marco Cavaliere and Pádraig Cantillon-Murphy. Enhancing electromagnetic tracking accuracy in medical applications using pre-trained witness sensor distortion models. *International Journal of Computer Assisted Radiology and Surgery*, 19(1):27–31, 2024.

[215] Alfred M Franz, Tamas Haidegger, Wolfgang Birkfellner, Kevin Cleary, Terry M Peters, and Lena Maier-Hein. Electromagnetic tracking in medicine—a review of technology, validation, and applications. *IEEE transactions on medical imaging*, 33(8):1702–1725, 2014.

[216] Gregorio Andria, Filippo Attivissimo, Attilio Di Nisio, Anna Maria Lucia Lanzolla, and Mattia Alessandro Ragolia. Analysis and optimization of surgical electromagnetic tracking systems by using magnetic field gradients. *Acta IMEKO*, 12(2):1–8, 2023.

[217] Nadja A Farshad-Amacker, Till Bay, Andrea B Rosskopf, José M Spirig, Florian Wanivenhaus, Christian WA Pfirrmann, and Mazda Farshad. Ultrasound-guided interventions with augmented reality in situ visualisation: a proof-of-mechanism phantom study. *European radiology experimental*, 4(1):1–7, 2020.

[218] Michele S Saruwatari, Trong N Nguyen, Hadi Fooladi Talari, Andrew J Matisoff, Karun V Sharma, Kelsey G Donoho, Sonali Basu, Pallavi Dwivedi,

James E Bost, and Raj Shekhar. Assessing the effect of augmented reality on procedural outcomes during ultrasound-guided vascular access. *Ultrasound in medicine & biology*, 49(11):2346–2353, 2023.

[219] Ho Chuen Kam, Ying Kin Yu, and Kin Hong Wong. An improvement on aruco marker for pose tracking using kalman filter. In *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 65–69. IEEE, 2018.

[220] Víctor Mondéjar-Guerra, Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Manuel J Marín-Jiménez, and Rafael Medina-Carnicer. Robust identification of fiducial markers in challenging conditions. *Expert Systems with Applications*, 93:336–345, 2018.

[221] Chang Peng, Qianqian Cai, Mengyue Chen, and Xiaoning Jiang. Recent advances in tracking devices for biomedical ultrasound imaging applications. *Micromachines*, 13(11):1855, 2022.

[222] Andrei State, Mark A Livingston, William F Garrett, Gentaro Hirota, Mary C Whitton, Etta D Pisano, and Henry Fuchs. Technologies for augmented reality systems: realizing ultrasound-guided needle biopsies. In *Proceedings of the 23rd annual conference on computer graphics and interactive techniques*, pages 439–446, 1996.

[223] D Magee, Y Zhu, Rish Ratnalingam, Peter Gardner, and David Kessel. An augmented reality simulator for ultrasound guided needle placement training. *Medical & biological engineering & computing*, 45(10):957–967, 2007.

[224] Marine Y Shao, Tamara Vagg, Matthias Seibold, and Mitchell Doughty. Towards a low-cost monitor-based augmented reality training platform for at-home ultrasound skill development. *Journal of Imaging*, 8(11):305, 2022.

[225] José N. Costa, João Gomes-Fonseca, Simão Valente, Luís Ferreira, Bruno Oliveira, Helena R. Torres, Pedro Morais, Victor Alves, and João L. Vilaça.

Ultrasound training simulator using augmented reality glasses: an accuracy and precision assessment study. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 4461–4464, 2022.

[226] C. Burden, J. Preshaw, P. White, T. J. Draycott, S. Grant, and R. Fox. Usability of virtual-reality simulation training in obstetric ultrasonography: a prospective cohort study. *Ultrasound in Obstetrics & Gynecology*, 42(2):213–217, 8 2013.

[227] Tobias Blum, Sandro Michael Heining, Oliver Kutter, and Nassir Navab. Advanced training methods using an augmented reality ultrasound simulator. In *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pages 177–178. IEEE, 2009.

[228] Nadja A Farshad-Amacker, Rahel A Kubik-Huch, Christoph Kolling, Cornelia Leo, and Jörg Goldhahn. Learning how to perform ultrasound-guided interventions with and without augmented reality visualization: a randomized study. *European Radiology*, 33(4):2927–2934, 2023.

[229] Maria Emine Nylund, Shubham Jain, Eva Tegnander, Eva Johanne Leknes Jensen, Ekaterina Prasolova-Førland, Frank Linsdeth, and Gabriel Kiss. Mixed reality training application to perform obstetric pulsed-wave doppler ultrasound. *Education and Information Technologies*, 29(6):7519–7551, 2024.

[230] Dehlela Shabir, Arshak Anjum, Hawa Hamza, Jhasketan Padhan, Abdulla Al-Ansari, Elias Yaacoub, Amr Mohammed, and Nikhil V Navkar. Development and evaluation of a mixed-reality tele-ultrasound system. *Ultrasound in Medicine & Biology*, 49(8):1867–1874, 2023.

[231] Qianhui Men, Clare Teng, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Multimodal-guidenet: Gaze-probe bidirectional guidance in obstetric

ultrasound scanning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 94–103. Springer, 2022.

[232] R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble. Automatic probe movement guidance for freehand obstetric ultrasound. In *MICCAI 2020: 23rd International Conference*. Springer, 2020.

[233] Yifan Cai, Richard Droste, Harshita Sharma, Pierre Chatelain, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. Spatio-temporal visual attention modelling of standard biometry plane-finding navigation. *Medical Image Analysis*, 65:101762, 2020.

[234] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[235] Yipei Wang, Qianye Yang, Lior Drukker, Aris Papageorghiou, Yipeng Hu, and J. Alison Noble. Task model-specific operator skill assessment in routine fetal ultrasound scanning. *International Journal of Computer Assisted Radiology and Surgery*, May 2022.

[236] Chiara Di Vece, Brian Dromey, Francisco Vasconcelos, Anna L David, Donald Peebles, and Danail Stoyanov. Deep learning-based plane pose regression in obstetric ultrasound. *International Journal of Computer Assisted Radiology and Surgery*, 17(5):833–839, 2022.

[237] Keyu Li, Jian Wang, Yangxin Xu, Hao Qin, Dongsheng Liu, Li Liu, and Max Q-H Meng. Autonomous navigation of an ultrasound probe towards standard scan planes with deep reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8302–8308. IEEE, 2021.

[238] Hannes Hase, Mohammad Farid Azampour, Maria Tirindelli, Magdalini Paschali, Walter Simson, Emad Fatemizadeh, and Nassir Navab. Ultrasound-guided robotic navigation with deep reinforcement learning. In *2020*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5534–5541. IEEE, 2020.

[239] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[240] J. Hein, M. Seibold, F. Bogo, M. Farshad, M. Pollefeys, P. Fürnstahl, and N. Navab. Towards markerless surgical tool and hand pose estimation. *International Journal of Computer Assisted Radiology and Surgery*, 16(5):799–808, 2021.

[241] R. Wang, S. Ktistakis, S. Zhang, M. Meboldt, and Q. Lohmeyer. Pov-surgery: A dataset for egocentric hand and tool pose estimation during surgical activities. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 440–450, 2023.

[242] E. D. Goodman, K. K. Patel, Y. Zhang, W. Locke, C. J. Kennedy, R. Mehrotra, S. Ren, M. Y. Guan, M. Downing, H. W. Chen, et al. A real-time spatiotemporal ai model analyzes skill in open surgical videos. *arXiv preprint arXiv:2112.07219*, 2021.

[243] A. Jin et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE WACV*, 2018.

[244] G. Lajkó, R. Nagyné Elek, and T. Haidegger. Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery. *Sensors*, 21, 2021.

[245] T. Nguyen, W. Plishker, A. Matisoff, K. Sharma, and R. Shekhar. Holous: Augmented reality visualization of live ultrasound images using hololens for ultrasound-guided procedures. *International Journal of Computer Assisted Radiology and Surgery*, 17, 2022.

[246] Brian P Dromey, Shahanaz Ahmed, Francisco Vasconcelos, Evangelos Mazomenos, Yada Kunpalin, Sebastien Ourselin, Jan Deprest, Anna L David, Danail Stoyanov, and Donald M Peebles. Dimensionless squared jerk: An objective differential to assess experienced and novice probe movement in obstetric ultrasound. *Prenatal Diagnosis*, page pd.5855, 11 2020.

[247] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[248] Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, and Derek Hoiem. Completing 3d object shape from one depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[249] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 536–551. Springer, 2014.

[250] Nasim Hajari, Gabriel Lugo Bustillo, Harsh Sharma, and Irene Cheng. Marker-less 3d object recognition and 6d pose estimation for homogeneous textureless objects: An rgb-d approach. *Sensors*, 20(18):5098, 2020.

[251] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[252] Mitchell Doughty and Nilesh R Ghugre. Hmd-egopose: Head-mounted display-based egocentric marker-less tool and hand pose estimation for augmented surgical guidance. *International journal of computer assisted radiology and surgery*, 17(12):2253–2262, 2022.

[253] Dorin Ungureanu, Federica Bogo, Silvano Galliani, Pooja Sama, Xin Duan, Casey Meekhof, Jan Stühmer, Thomas J Cashman, Bugra Tekin, Johannes L Schönberger, et al. Hololens 2 research mode as a tool for computer vision research. *arXiv preprint arXiv:2008.11239*, 2020.

[254] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[255] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[256] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[257] Jonas Hein, Nicola Cavalcanti, Daniel Suter, Lukas Zingg, Fabio Carrillo, Lilian Calvet, Mazda Farshad, Marc Pollefeys, Nassir Navab, and Philipp Fürnstahl. Next-generation surgical navigation: Marker-less multi-view 6dof pose estimation of surgical instruments. *arXiv preprint arXiv:2305.03535*, 2023.

[258] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, November 2017.

[259] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF CVPR*, June 2020.

[260] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan

Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021.

[261] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics Automation Magazine*, 22(3):36–52, 2015.

[262] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021.

[263] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF CVPR*, pages 14687–14697, June 2021.

[264] Theocharis Chatzis, Andreas Stergioulas, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. A comprehensive study on deep learning-based 3d hand pose estimation methods. *Applied Sciences*, 10(19):6850, 2020.

[265] Joseph HR Isaac, Muniyandi Manivannan, and Balaraman Ravindran. Single shot corrective cnn for anatomically correct 3d hand pose estimation. *Frontiers in Artificial Intelligence*, 5:759255, 2022.

[266] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

[267] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *European Conference on Computer Vision*, pages 183–202. Springer, 2024.

[268] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

[269] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: `https://grab.is.tue.mpg.de`.

[270] Haoming Li, Xinzhuo Lin, Yang Zhou, Xiang Li, Yuchi Huo, Jiming Chen, and Qi Ye. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. *arXiv preprint arXiv:2210.09245*, 2022.

[271] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015.

[272] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Ho-3d: A multi-user, multi-object dataset for joint 3d hand-object pose estimation. *CoRR*, 2019.

[273] B. Doosti, S. Naha, M. Mirbagheri, and D. Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*, June 2020.

[274] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF ICCV*, 2019.

[275] Andrew T. Miller and Peter K. Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004.

[276] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245, November 2017.

[277] Matei T Ciocarlie and Peter K Allen. Hand posture subspaces for dexterous robotic grasping. *The International Journal of Robotics Research*, 28(7):851–867, 2009.

[278] Blender Online Community. Blender - a 3d modelling and rendering package. `http://www.blender.org`, 2018.

[279] Abdulkadir Akin, E. Erdede, Hossein Afshari, Alexandre Schmid, and Yusuf Leblebici. Enhanced omnidirectional image reconstruction algorithm and its real-time hardware. In *Proceedings - 15th Euromicro Conference on Digital System Design, DSD 2012*, September 2012.

[280] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 361–378, Cham, 2020. Springer International Publishing.

[281] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Tech Report*, 2015.

[282] H. Gao and S. Ji. Graph u-nets. In *ICML*, 2019.

[283] Hongyang Gao and Shuiwang Ji. Graph u-nets. In *international conference on machine learning*, pages 2083–2092. PMLR, 2019.

[284] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina

Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2636–2645, 2018.

[285] Georgios Paschalidis, Romana Wilschut, Dimitrije Antić, Omid Taheri, and Dimitrios Tzionas. 3d whole-body grasp synthesis with directional controllability. *arXiv preprint arXiv:2408.16770*, 2024.

[286] Purva Tendulkar, Dídac Surís, and Carl Vondrick. Flex: Full-body grasping without full-body grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21179–21189, 2023.

[287] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Wendelin Knauer, Klaus H Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023.

[288] Alice HY Chan, Wei Feng Lee, Pascal WM Van Gerven, and Jordan Chenkin. Assessment of changes in gaze patterns during training in point-of-care ultrasound. *BMC Medical Education*, 22(1):658, 2022.

[289] KJ Rittenhouse, B Vwalika, Y Sebastiao, T Pokaprakarn, N Sindano, H Shah, EM Stringer, MP Kasaro, SR Cole, JSA Stringer, et al. Accuracy of portable ultrasound machines for obstetric biometry. *Ultrasound in Obstetrics & Gynecology*, 63(6):772–780, 2024.

[290] Bryan J Ranger, Elizabeth Bradburn, Qingchao Chen, Micah Kim, J Alison Noble, and Aris T Papageorghiou. Portable ultrasound devices for obstetric care in resource-constrained environments: mapping the landscape. *Gates Open Research*, 7:133, 2024.

[291] Xingyu Liu, Pengfei Ren, Yuanyuan Gao, Jingyu Wang, Haifeng Sun, Qi Qi, Zirui Zhuang, and Jianxin Liao. Keypoint fusion for rgb-d based 3d hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3756–3764, 2024.

[292] Qiuxia Lin, Linlin Yang, and Angela Yao. Cross-domain 3d hand pose estimation with dual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17184–17193, 2023.

[293] Vanessa Wirth, Anna-Maria Liphardt, Birte Coppers, Johanna Bräunig, Simon Heinrich, Sigrid Leyendecker, Arnd Kleyer, Georg Schett, Martin Vossiek, Bernhard Egger, and Marc Stamminger. Sharpy: Shape reconstruction and hand pose estimation from rgb-d with uncertainty. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2625–2633, October 2023.

[294] Jemin Hwangbo, Joonho Lee, and Marco Hutter. Per-contact iteration method for solving contact dynamics. *IEEE Robotics and Automation Letters*, 3(2):895–902, 2018.

[295] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022.

[296] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. GraspXL: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision (ECCV)*, 2024.

[297] Linyi Huang, Hui Zhang, Zijian Wu, Sammy Christen, and Jie Song. Fungrasp: Functional grasping for diverse dexterous hands. *arXiv preprint arXiv:2411.16755*, 2024.

[298] Joonho Lee, Jemin Hwangbo, and Marco Hutter. Robust recovery controller for a quadrupedal robot using deep reinforcement learning. *arXiv preprint arXiv:1901.07517*, 2019.

[299] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

[300] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):eabc5986, 2020.

[301] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, 2023.

[302] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*, pages 235–246. IEEE, 2024.

[303] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[304] Silvia Zaccardi, Redona Brahimetaj, Federico Trovalusci, Reinhard Claeys, Rossana Lovecchio, David Beckwée, Eva Swinnen, and Bart Jansen. Insights into azure kinect skeletal tracking: A simple approach to reduce ir passive noise. In *2025 25th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2025.

[305] Adiba Orzikulova, Diana A Vasile, Chi Ian Tang, Fahim Kawsar, Sung-Ju Lee, and Chulhong Min. Bioq: Towards context-aware multi-device collaboration with bio-cues. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 504–517, 2025.

[306] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023.

[307] Jian Yang, Jianjun Zhu, Daniel Y Sze, Li Cui, Xiaohui Li, Yanhua Bai, Danni Ai, Jingfan Fan, Hong Song, and Feng Duan. Feasibility of augmented reality–guided transjugular intrahepatic portosystemic shunt. *Journal of Vascular and Interventional Radiology*, 31(12):2098–2103, 2020.