Medical Image Analysis 71 (2021) 102053

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

CaDIS: Cataract dataset for surgical RGB-image segmentation

Maria Grammatikopoulou[a,1,*], Evangello Flouty[a,1], Abdolrahim Kadkhodamohammadi[a], Gwenolé Quellec[b], Andre Chow[a], Jean Nehme[a], Imanol Luengo[a], Danail Stoyanov[a,c]

[a] Digital Surgery LTD, 230 City Road, London, EC1V 2QY, UK
[b] Inserm, UMR 1101, Brest F-29200, France
[c] Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, Gower Street, London, WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 28 May 2020
Revised 17 March 2021
Accepted 23 March 2021
Available online 31 March 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Cataract surgery
Semantic segmentation
Dataset

ABSTRACT

Video feedback provides a wealth of information about surgical procedures and is the main sensory cue for surgeons. Scene understanding is crucial to computer assisted interventions (CAI) and to post-operative analysis of the surgical procedure. A fundamental building block of such capabilities is the identification and localization of surgical instruments and anatomical structures through semantic segmentation. Deep learning has advanced semantic segmentation techniques in the recent years but is inherently reliant on the availability of labelled datasets for model training. This paper introduces a dataset for semantic segmentation of cataract surgery videos complementing the publicly available CATARACTS challenge dataset. In addition, we benchmark the performance of several state-of-the-art deep learning models for semantic segmentation on the presented dataset. The dataset is publicly available at https://cataracts-semantic-segmentation2020.grand-challenge.org/.

1. Introduction

Computer assisted interventions (CAI) have the potential to enhance surgeons' capabilities through better clinical information fusion, navigation and visualization (Maier-Hein et al., 2017). Currently, CAI systems are used mainly as tools for pre-operative planning (Zeng et al., 2016). There is potential to develop CAI further for improved navigation capabilities, better imaging and robotic instrumentation (Kassahun et al., 2016). However, such systems depend on effective use of surgical video.

Data-driven machine learning techniques and deep learning have been immensely influential in recent computer vision advances as well as in medical image computing and analysis. Therefore, using surgical cameras, establishing data repositories and data labelling to facilitate training of vision models and subsequent benchmarking is necessary to exploit such advances for CAI (Maier-Hein et al., 2017; Vedula et al., 2017).

Pixel-level annotations are necessary for model development, and in particular for image segmentation models. Such models could advance applications such as image-guided interventions (Kovler et al., 2015; Pfeiffer et al., 2019), support pre-operative surgical planning (Ozdemir and Goksel, 2019), estimate instrument usage and motion for post-operative analytics (Allan, 2015; Fuentes-Hurtado, 2019; García-Peraza-Herrera, 2016), automate diagnostic readouts (Bouget, 2019; Suetens, 1993) and enhance surgical training (Engelhardt et al., 2018). While data availability is increasingly growing through the usage of digital surgical cameras in endoscopy, laparoscopy and microsurgery, and due to well-established systems for managing confidentiality, regulation and ethics, annotation and data labelling are still a major challenge for CAI.

Over the past decade, the emergence of surgical video datasets has significantly contributed to the progress of computer vision-based CAI systems. Notable examples include the Cholec80, Cholec120 (Twinanda et al., 2016), RMIT (Sznitman et al., 2012) and the EndoVis challenge datasets.[2] In particular, two robotic instrument segmentation datasets have been released for the 2017 (Allan et al., 2019) and 2018 (Allan et al., 2020) Robotic Instrument Segmentation EndoVis sub-challenges that included segmentation masks for robotic instruments appearing in the scene. The 2017 Robotic Instrument Segmentation dataset was later extended

[*] Corresponding author.
E-mail address: maria.grammatikopoulou@medtronic.com
(M. Grammatikopoulou).

[1] M. Grammatikopoulou and E. Flouty contributed equally to this work.

[2] https://endovis.grand-challenge.org/

https://doi.org/10.1016/j.media.2021.102053

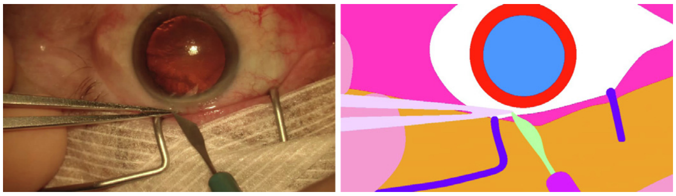


Fig. 1. Example image frame (left) and semantic segmentation labels (right) from the Cataract dataset for Image Segmentation presented in this paper. (Colormap: 🟦 Pupil, 🟥 Iris, ⬜ Cornea, 🟪 Skin, 🟧 Surgical tape, 🟦 Eye retractors, 🟪 Hand, ⬜ Bonn Forceps, 🟩 Secondary Knife and 🟪 Secondary Knife Handle).

for the 2018 Robotic Scene Segmentation EndoVis sub-challenge to include pixel-wise labels for anatomical structures for approximately 2400 endoscopic images (Allan et al., 2020). While releasing these datasets, the research community has also worked towards standardizing the reporting of datasets and challenges (Maier-Hein et al., 2020).

Recently, the CATARACTS challenge[3] presented 50 annotated surgical videos obtained through a surgical microscope (Al Hajj et al., 2019). The dataset was annotated to provide both frame-level instrument presence labels and frame-level surgical phase labels (Al Hajj et al., 2019; Zisimopoulos et al., 2018). From a clinical perspective, even though cataract surgery is less prone to complications, risk mitigation can have big impact, with over 20 million cases recorded in 2010 (WHO, 2018). Studies on medical malpractice related to cataract surgery revealed that 76.28% of the 118 claims are intra-operative allegations (Kim et al., 2012) and that the rate of a certain intra-operative complication (posterior capsular rent) was 0.45–3.6% for experienced surgeons, and 0.8–8.9% for residents (Chakrabarti and Nazm, 2017). With these in consideration, a dataset for semantic segmentation may lead to the development of systems that could potentially reduce complications and improve workflow.

In this paper, we introduce a semantic segmentation dataset generated from videos of the training set of the CATARACTS dataset. The dataset includes pixel-wise annotations for the entire surgical scene for cataract surgery procedures, including anatomical structures and surgical instruments, for 4670 surgical microscope images (Fig. 1). The aim of releasing such a dataset is to allow simultaneous anatomy and instrument pixel-level localization. A potential application is the detection of anatomy and surgical instrument interactions, which can be subsequently used to assess the safety and progress of the surgical procedure. We demonstrate how this dataset can be used to train state-of-the-art deep learning frameworks to segment microscope images from cataract surgery. We believe this contribution will underpin the development of CAI techniques based on surgical vision.

2. Cataract dataset for image segmentation

The dataset was generated from the training videos released for the CATARACTS challenge (Al Hajj et al., 2019). The CATARACTS challenge training set includes 25 videos with average duration of 10 min and 56 s recorded at 30 frames per second (fps).

2.1. Data sources

The recorded operations were performed in Brest University Hospital from January to September 2015 (Al Hajj et al., 2019). The videos were recorded using a 180I camera (Toshiba, Tokyo, Japan) mounted on an OPMI Lumera T microscope (Carl Zeiss Meditec, Jena, Germany) focusing on the patient's eye. The surgeries were performed by three surgeons of varying expertise levels (one expert, one mid-level and an intern surgeon). The average age of the patients was 61 years old, with a minimum of 23, a maximum of 83 years old and 10 years standard deviation. The surgeries were performed because of age-related causes, trauma and refractive errors. Each video corresponds to a different patient. The study was approved by the Institutional Review Board of Brest University Hospital on 28 January 2013. All patients were informed and gave their consent to participate in the study.

Table 1
Phases sampled per video in CaDIS dataset. Phase numbering in the table as defined in (Zisimopoulos et al., 2018) The defined phases are: (1) Access of anterior chamber: sideport incision, (2) Access of anterior chamber: mainport incision, (3) Lens removal: Viscoelastic injection, (4) Lens removal, (5) Phacoemulsification: Viscoelastic injection, (6) Phacoemulsification: Capsulorhexis, (7) Phacoemulsification: Lens hydrodissection, (8) Phacoemulsification, (9) Phacoemulsification: Lens matter removal, (10) Lens insertion: Viscoelastic injection, (11) Lens insertion, (12) Aspiration of viscoelastic, (13) Wound closure and (14) Wound closure with suture.

Video IDs	Phases sampled in video
Video 1	1, 3, 7, 8, 9, 10, 11
Video 2	1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12
Videos 3, 5	1, 2, 3, 5, 6, 7, 8, 9, 10
Video 4	1, 2, 3, 4, 5, 6, 7, 8, 9
Video 6	7, 8, 9, 10, 11, 12
Videos 7, 9, 11	1, 2, 3, 5, 6, 7, 8, 9
Video 8, 10, 15-18	1, 2, 3, 5, 6, 7, 8
Video 12	1, 2, 3, 4, 6, 7
Video 13	1, 2, 3, 5, 6, 7
Video 14	1, 2, 3, 4, 5, 6
Videos 19, 23, 24	1, 2, 3, 5, 7, 8, 9
Videos 20, 21	1, 2, 3, 5, 6, 7, 8, 9, 10
Video 22	1, 2, 4, 5, 7, 8, 9
Video 25	1, 2, 3, 5, 7, 8, 9, 10, 11

2.2. Training and test set characteristics

Frames from the 25 training videos were extracted using the reference instrument and phase information. This is to select video frames that include instruments and to ensure a class distribution across the surgical phases that represents real-world scenarios. In particular, the videos were sampled to tackle the overhead of pixel-level manual labelling for semantic mask generation in order to label as many frames, which contain substantial scene variations. The surgical procedures were divided into 14 phases as in (Zisimopoulos et al., 2018). The phases sampled per video in the presented dataset are given in Table 1. A number of 10 to 20 frames were randomly selected per phase such that the frames are at least 0.3 s apart. The images were also resized from 1920 × 1080 to 960 × 540. In total, 4670 frames were selected.

2.3. Annotation process

After frame selection, the videos were annotated manually. The guidelines for anatomy and instrument annotation were drafted by a team of in-house expert medical officers. A team of four in-house roto artists (annotators) created the pixel-wise segmentation masks. The annotators used commercial rotoscoping software to create the segmentation masks. The annotators were trained by the medical officers in order to get familiar with the surgical procedure, the anatomical structures appearing and the different instruments used at each phase. The annotators had direct access to the medical officers at all steps of the annotation process. Every frame was annotated by one roto artist. To ensure the quality of annotations, every annotated frame was also checked by a second annotator. In case of disagreement between the annotators, the medical officers' opinion is sought in accordance with the

[3] https://cataracts.grand-challenge.org/

specified annotation guidelines. The medical officers validated the segmentation mask annotations. Further pixel-wise checks per segmentation mask were performed by programmatically extracting all contours from the generated segmentation masks and overlaying them to the respective image frame. This facilitated visual inspection of the segmentation masks to ensure accurate anatomy and instrument boundaries. In addition, pixel-wise checks were performed to ensure that all clusters of pixel larger than 50 pixels are assigned to a class. The same process of annotation was applied to all selected frames (training, validation and test set).

2.4. Sources of error

Potential sources of error in the annotation can be attributed to blurriness due to substantial instrument or patient motion. This contributes to having instrument or anatomy out of focus and, therefore, not have very clear boundaries in some frames. However, even in this cases, it was ensured that the instrument and anatomy boundaries are as accurate as possible. Specular reflections may also lead to inaccurate boundary delineation, especially for the instrument tips when they are inside the anatomy.

2.5. Dataset statistics

The dataset includes 36 classes: 29 surgical instrument classes, 4 anatomy classes and 3 miscellaneous classes. The list of classes per category and the statistics of the dataset are given in Table 2. As expected, the anatomy classes appear more frequently than the surgical instruments. The anatomy also covers the largest part of the scene, as it can be seen from the average number of pixels that represent the pupil, iris and cornea compared to the surgical instruments (Table 2). In addition, the Presence In Videos metric shows that 17 instrument classes appear in less than half of the videos. The instance and pixel distribution indicate that the dataset is imbalanced and, consequently, accurate instrument classification is more challenging. Furthermore, there are other visual challenges due to the high inter-class similarity among instruments. For example, Fig. 2 shows four different types of cannulas, which look very similar. Each of these cannulas are used to perform different actions, like injecting material and handling tissue. Therefore, as the type of instrument can reveal information and be one of the main indications of what surgical action has been performed, it is of interest to distinguish different instrument types.

3. Experiments

A set of experiments were performed using the presented dataset in three different tasks as described in the following sections. Baseline experiments were performed using state-of-the-art segmentation networks (Table 3) to provide a reference for future experiments using the dataset. It is worth noting that the models were not optimized to each given task.

3.1. Tasks

Three tasks are presented that use different class groupings. The motivation for the following tasks is that anatomical structure and instrument localization could be useful for intra- and post-operative image guidance and risk assessment. Instrument segmentation and identification can be useful to a different degree, for example identifying only where anatomy or an instrument is or additionally identifying the types of instruments. A brief description of each task is given in the following sections.

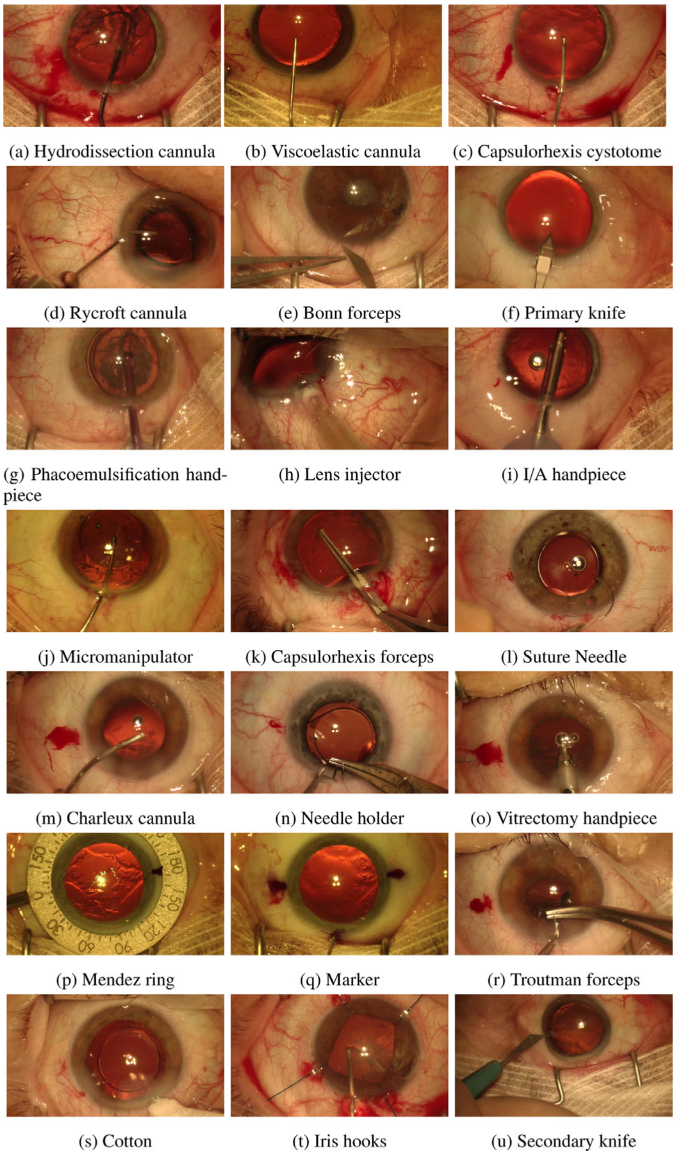

(a) Hydrodissection cannula (b) Viscoelastic cannula (c) Capsulorhexis cystotome
(d) Rycroft cannula (e) Bonn forceps (f) Primary knife
(g) Phacoemulsification handpiece (h) Lens injector (i) I/A handpiece
(j) Micromanipulator (k) Capsulorhexis forceps (l) Suture Needle
(m) Charleux cannula (n) Needle holder (o) Vitrectomy handpiece
(p) Mendez ring (q) Marker (r) Troutman forceps
(s) Cotton (t) Iris hooks (u) Secondary knife

Fig. 2. Instances for all instruments appearing in the dataset.

3.1.1. Task I

The first task is focused on differentiating between anatomy and instruments within every frame. Therefore, instrument classification is excluded from this task as the purpose of this scenario is to identify mainly where the anatomical structures are. This task is defined by segmenting the scene into 8 classes, and in particular, 4 classes for anatomical structures, 1 class for all instruments and 3 classes for all other objects appearing in the scene (Table 5).

3.1.2. Task II

The second task incorporates instrument classification and includes 17 classes given in Table 7. The instruments are grouped in categories according to appearance similarities and instrument types as suggested by the medical officers who created the annotation guidelines. This task is to identify anatomical structures and also the main types of instruments that appear in the scene. The purpose of identifying the main instrument type simultaneously is to give more information on the stage of the procedure through scene segmentation. The type of instrument can also help differentiating overlapping instruments in the segmentation output, in case of use of different instruments, which would otherwise

Table 2
Total instances per class [% of total number of frames in each set], total presence per class in videos [% of total videos in each set] and average number of pixels per class per frame for the entire dataset and for the training, validation and test splits.

Category	ID class name	All videos			Training set		Validation set		Test set	
		Instances per class	Presence in videos	Avg. pixels per class	Instances per class	Presence in videos	Instances per class	Presence in videos	Instances per class	Presence in videos
Anatomy	0 Pupil	99.9	100	87,215	99.8	100	100	100	100	100
	4 Iris	99.9	100	58,247	99.9	100	100	100	100	100
	5 Skin	99.9	100	69,351	100	100	98.9	100	100	100
	6 Cornea	100	100	253,631	100	100	100	100	100	100
Instruments	7 Hydrosdissection Cannula	9.6	100	6840	9.6	100	9.7	100	9.2	100
	8 Viscoelastic Cannula	12.6	100	3697	13	100	10.9	3	11.4	100
	9 Capsulorhexis Cystotome	9.6	100	5016	9.4	100	10.5	100	9.9	100
	10 Rycroft Cannula	9.4	100	3571	9.2	100	10.1	100	10.2	100
	11 Bonn Forceps	8.2	88	16,476	8	84.2	5.1	100	12.6	100
	12 Primary Knife	6.6	96	11,040	6.7	94.7	5.6	100	7	100
	13 Phacoemulsifier Handpiece	9.8	100	9745	9.6	100	11	3	10.1	3
	14 Lens Injector	8.6	96	19,543	8.2	94.7	10.3	100	9.9	100
	15 Irrigation/Aspiration (I/A) Handpiece	16.6	92	11,291	15.9	89.4	21	100	16.4	100
	16 Secondary Knife	6.4	100	8644	6.4	100	5.8	100	6.5	100
	17 Micromanipulator	13.3	100	7690	13	100	15.2	100	13.5	100
	18 Irrigation/Aspiration Handpiece Handle	2.1	68	12,894	1.9	57.9	1.3	100	4.6	100
	19 Capsulorhexis Forceps	2.8	48	13,268	3	47.4	1.5	66.7	2.4	33.3
	20 Rycroft Cannula Handle	1.8	52	10,556	1.5	42.1	3.4	100	2.4	66.7
	21 Phacoemulsifier Handpiece Handle	1.5	40	16,199	1.6	42.1	1.3	33.3	1.2	33.3
	22 Capsulorhexis Cystotome Handle	1.8	44	4993	1.7	36.8	2.1	33.3	2.2	100
	23 Secondary Knife Handle	2.8	80	10,004	3	78.9	2.2	66.7	2.6	100
	24 Lens Injector Handle	0.9	16	17,670	0.5	10.5	2.4	33.3	1.4	33.3
	25 Suture Needle	0.7	16	802	0.7	15.8	0	0	1.2	33.3
	26 Needle Holder	0.3	4	31,156	0.3	5.3	0	0	0	0
	27 Charleux Cannula	0.4	8	5042	0.6	10.5	0	0	0	0
	28 Primary Knife Handle	0.1	8	2395	0.01	5.3	0	0	0.3	33.3
	29 Vitrectomy Handpiece	0.4	4	14,637	0.5	5.3	0	0	0	0
	30 Mendez Ring	0.1	4	151,711	0.2	5.3	0	0	0	0
	31 Marker	3.6	4	7034	4.8	5.3	0	0	0	0
	32 Hydrosdissection Cannula Handle	0.3	8	2291	0.3	10.5	0	0	0	0
	33 Troutman Forceps	0.4	8	22,246	0.2	5.3	0	0	2.4	33.3
	34 Cotton	0.4	12	16,623	0.6	15.8	0	0	0	0
	35 Iris Hooks	2.7	4	4525	3.5	5.3	0	0	0	0
Others	1 Surgical Tape	77	96	39,907	72	94.7	86.7	100	98.5	100
	2 Hand	13	100	29,473	12.7	100	10.3	100	17.2	100
	3 Eye Retractors	73.5	100	4033	71.7	100	93.4	100	66.6	100

Table 3
Number of parameters of baseline models.

Model	Number of parameters
UNet	14 M
DeepLabV3 (Xception/Mobilenet)	58 M / 54 M
UPerNet	60 M
HRNetV2	66 M

Table 4
Mean Intersection over Union (mIoU) [%], Pixel Accuracy (PA) [%] and Pixel Accuracy per Class (PAC) [%] per model for validation and test sets for Task I.

Model	Validation set			Test set		
	mIoU	PA	PAC	mIoU	PA	PAC
UNet	86.7	94.8	92.3	83.7	94.3	89.3
DeepLabV3+	85.3	94.8	92	82.6	93.9	88.7
UPerNet	87.9	95.4	93.2	84	94.2	89.5
HRNetV2	88.1	95.2	93	84.9	94.2	90

Table 5
mIoU per class [%] for test set for Task I.

	UNet	DeepLabV3+	UPerNet	HRNetV2
Pupil	94	93.9	93.9	94.2
Surgical Tape	86.9	84.4	87	88.5
Hand	85.2	82.4	86	86.5
Eye Retractors	80.2	81.7	82.6	87.5
Iris	84.9	84.5	84.6	85
Skin	70.9	67	68.1	68
Cornea	93.2	92.8	93	92.7
Instrument	73.8	74.3	76.4	77
mIoU (Anatomy)	85.8	84.5	84.9	85
mIoU (Instruments)	73.8	74.3	76.4	77
mIoU (All classes)	83.7	82.6	84	84.9

Table 6
Mean Intersection over Union (mIoU) [%], Pixel Accuracy (PA) [%] and Pixel Accuracy per Class (PAC) [%] per model for validation and test sets for Task II.

Model	Validation set			Test set		
	mIoU	PA	PAC	mIoU	PA	PAC
UNet	72.7	94.9	82.8	70.6	94	79.6
DeepLabV3+	74.5	94.4	83.3	72.3	93.5	80.8
UPerNet	79.5	95	86.8	73.8	94.1	82
HRNetV2	81.8	95.4	88.6	76.1	94.6	83.6

Table 7
mIoU per class [%] for test set for Task II.

	UNet	DeepLabV3+	UPerNet	HRNetV2
Pupil	93.8	94	94	94
Surgical Tape	85.3	82.9	87.3	90
Hand	84.6	83.8	86	86.7
Eye Retractors	79.8	80.6	86	86.5
Iris	84.9	84.4	84.9	85
Skin	69.5	64.8	67.8	72.6
Cornea	93	92.4	93	93.4
Cannula	44.5	48.9	50	49.5
Cap. Cystotome	40.4	55.7	54.5	61.7
Tissue Forceps	65	70	74	78
Primary Knife	87	86.1	89.5	89.3
Ph. Handpiece	74.7	75	77.6	77.9
Lens Injector	79	78.5	81	82.8
I/A Handpiece	69.5	74	73.6	75.3
Secondary Knife	74.7	69	68.2	79.5
Micromanipulator	51.4	59.3	63.6	64.4
Cap. Forceps	22.9	28.9	23	27.2
mIoU (Anatomy)	85.4	83.9	84.9	86.3
mIoU (Instruments)	60.9	64.6	65.5	68.6
mIoU (All classes)	70.6	72.3	73.8	76.1

Table 8
Mean Intersection over Union (mIoU) [%], Pixel Accuracy (PA) [%] and Pixel Accuracy per Class (PAC) [%] per model for validation and test sets for Task III.

Model	Validation set			Test set		
	mIoU	PA	PAC	mIoU	PA	PAC
UNet	66.6	94.7	78.9	59.2	93.9	70.5
DeepLabV3+	68.6	94.5	79.9	63.2	93.9	75.6
UPerNet	74.2	95.3	84.7	66.8	94.2	77.8
HRNetV2	72.4	95.3	83.1	66.6	94.3	77

Table 9
mIoU per class [%] for test set for Task III.

	UNet	DeepLabV3+	UPerNet	HRNetV2
Pupil	94	93.9	94	94.1
Surgical Tape	87.2	87.1	87.4	88.9
Hand	84.5	82.3	85.3	86.4
Eye Retractors	83.8	82.8	84.2	87.3
Iris	84.4	84.3	85.1	84.6
Skin	68.9	68.7	68.5	70
Cornea	92.9	92.5	93.2	93.1
Hydro. Cannula	45.6	53.7	54.6	55.2
Visc. Cannula	39.5	57	57.4	62.7
Cap. Cystotome	42.1	41.4	58.3	60.6
Rycroft Cannula	40.8	52	54.5	56.2
Bonn Forceps	70.7	66.9	76.8	77.2
Primary Knife	84.1	87.9	90.6	90.5
Ph. Handpiece	75.7	74.8	77.5	77
Lens Injector	69.8	72.6	72.9	71.1
I/A Handpiece	69.5	71.8	71.7	72.9
Secondary Knife	77.1	79.4	88.6	89.5
Micromanipulator	55.4	59.7	61.1	64.6
I/A Handpiece Handle	67.8	72.1	74.5	71.5
Cap. Forceps	19.7	33.7	40.1	36.3
R. Cannula Handle	9.8	23.3	33	20
Ph. Handpiece Handle	56.9	46.8	65.4	60.4
Cap. Cystotome Handle	28.7	67.7	64.6	54
Sec. Knife Handle	30.4	29	29.9	42.1
Lens Injector Handle	0	0	0	0
mIoU (Anatomy)	85.3	84.8	85.6	86.2
mIoU (Instruments)	52	58.2	63	62.5
mIoU (All classes)	59.2	63.2	66.8	66.6

be shown as one merged area. In addition, grouping instrument classes mitigates class imbalance while also allows a degree of instrument classification in combination with anatomy segmentation. The classes that are merged are: (i) hydrosdissection cannula and handle, viscoelastic cannula, Rycroft cannula and handle and Charleux cannula as cannula and (ii) Bonn and Troutman forceps as tissue forceps while all the other instrument classes were merged with their respective handle. It should be noted that the instruments that appear only in the training set and could not be merged with another instrument class were ignored during training. The ignored classes are: suture needle, needle holder, vitrectomy handpiece, marker, cotton, iris hooks and Mendez ring.

3.1.3. Task III

The third task includes 25 classes as listed in Table 9. This task allows more granular instrument classification by representing each instrument type and instrument tips and handles as separate classes. The classes that do not appear in all splits and are present in less than 5 videos were ignored during training (Table 2). Identifying all instrument types in the scene gives even more explicit information about the stage of the surgery. For example, different cannulas are used in different phases of the procedure. In addition, segmenting instrument tips and handles can result to more accurate information of which part of the instrument interacts with the anatomy. Therefore, the third tasks aims at combining anatomy and instrument segmentation while giving the most information about the procedure itself through identifying the exact type and part of the instrument.

3.2. Dataset splits

The videos are separated into training, validation and test sets. The dataset distribution per set is presented in Table 2. As not all classes are present in all videos, we ensured that the videos in the training set include samples from all classes. We split the rest of the videos between the validation and test sets so that sufficient instrument instances were present in each set to allow a fair assessment of models across different instrument classes. It should be mentioned that, since the split is done on a video basis and not all instruments appear with the same proportion in all videos, that there is an inherent difference in class distribution among the three splits. The distribution of classes in splits could also be done on a frame basis for a more uniform class distribution between training, validation and testing. However, dividing frames from the same video across training, validation and test sets was avoided as it would result in training and evaluating the model on frames from previously seen videos. The training, validation and test sets contain 3550, 534 (Videos 5, 7 and 16) and 586 (Videos 2, 12 and 22) images respectively. As mentioned in Section 2.5, the dataset is imbalanced since the classes that represent instruments appear less frequently and occupy less pixels per frame than the anatomy (Table 2). The main problems that are present with imbalance are that it is more challenging to learn and classify less frequent instruments or smaller instruments.

3.3. Metrics

The metrics that are used to assess the segmentation quality are the mean Intersection over Union (mIoU), Pixel Accuracy (PA) and Pixel Accuracy per Class (PAC) and the IoU per class. The formulations for PA, PAC and mIoU are defined as follows:

$$\mathrm{PA} = \frac{\sum_{i=1}^{N} p_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij}}, \quad i,j = 1,\ldots,N \tag{1}$$

$$\mathrm{PAC} = \frac{1}{N}\sum_{i=1}^{N}\frac{p_{ii}}{\sum_{j=1}^{N} p_{ij}} \quad i,j = 1,\ldots,N \tag{2}$$

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{p_{ii}}{\sum_{j=1}^{N} p_{ij} - p_{ii} + \sum_{j=1}^{N} p_{ji}}, \quad i,j = 1,\ldots,N \tag{3}$$

were N the number of classes and p_{ij} the number of pixels predicted as class i and labelled as class j. It is worth noting that the ignored classes were not taken into account when the metrics are calculated.

3.4. Baseline models

The three tasks are benchmarked on state-of-the-art models to provide a baseline for semantic segmentation models for cataract surgery. The models used in the baseline experiments are UNet (Ronneberger et al., 2015), DeepLabV3+ (Chen et al., 2018), UPerNet (Xiao et al., 2018) and HRNetV2 (Wang et al., 2020). UNet was proposed by Ronnenberger et al. for biomedical image segmentation. It has been widely used in the medical community because of its relatively low number of parameters. DeepLabV3+ was introduced as an extension of DeepLab (v2 (Chen et al., 2017a) and v3 (Chen et al., 2017b)) that uses modified Xception (Chollet, 2017) as the encoder and combines it with atrous convolutions with different dilation rates to achieve better contextual predictions without losing image resolution. The atrous convolution enables DeepLabV3+ to benefit from long-range contextual information while preserving fine boundary information. In this work, MobilenetV2 (Sandler et al., 2018) is used as the backbone for DeepLabV3+ in order to use a light-weight version of the model. UperNet uses a pyramid pooling module to make use of both global and local contextual information. To extract and incorporate this information, the model relies on a Feature Pyramid Network to extract features at different scales of the encoder, which allows to build a richer representation by combining information at multiple image scales. Lastly, HRNetV2 attempts to preserve high-resolution feature representations by combining features from all scales throughout the encoder and also from parallel convolution streams. The open-source implementations of the networks were used in all experiments (UNet,[4] DeepLabV3+,[5] UperNet,[6] HRNetV2[7]).

3.5. Training process

3.5.1. Data pre/post-processing

Data augmentation was applied prior to model training. The same augmentation was applied for all models. Each training image was normalized, flipped, randomly rotated and hue and saturation was also adjusted. The input images were downsized to 270 × 480. No post-processing was performed.

3.5.2. Experiment parameters and setup

The network weights for UPerNet and HRNetV2 were initialized using pre-trained weights on ImageNet (Deng et al., 2009) while for DeepLabV3+ pre-trained weights on Pascal VOC (Everingham et al., 2010) were used. The networks were trained on a system with two NVIDIA GTX 1080 Ti GPUs for 100 epochs. For all models, the Cross Entropy loss function was used with learning rate equal to 10^{-4} using the Adam Optimizer. The β_1, β_2 and ϵ values for the Adam Optimizer were set to 0.9, 0.999 and 10^{-8}, which are proposed as good default values for the optimizer in (Kingma and Ba, 2014). It is noted that default parameters which are fair for all models were chosen, rather than optimizing each model's hyperparameters separately.

3.6. Results

3.6.1. Task I

The overall mIoU, PA and PAC for the validation and test set for all models in Task I are given in Table 4. In particular, for anatomy segmentation of the test set, UNet presents a mIoU of 85.8 %, DeepLabV3+ of 84.5%, UPerNet of 84.9% and HRNetV2 of 85% (Table 5). Similarly for instrument segmentation, UNet gives a mIoU of 73.8%, DeepLabV3+ 74.3%, UPerNet of 76.4% and HRNetV2 of 77% (Table 5).

3.6.2. Task II

The mIoU, PA and PAC for the validation and test set are shown in Table 6. The mIoUs for anatomy segmentation are 85.4%, 83.9%, 84.9% and 86.3% for UNet, DeepLabV3+, UPerNet and HRNetV2 respectively (Table 7). For instrument segmentation for Task II, the IoUs per class are 60.9%, 64.6%, 65.5% and 68.6% for UNet, DeepLabV3+, UPerNet and HRNetV2 (Table 7).

3.6.3. Task III

The results for Task III for the validation and test set are given in Table 8. Including now all tips and instrument handles as separate classes, for instrument segmentation the mIoUs for UNet,

[4] UNet: https://github.com/milesial/Pytorch-UNet

[5] DeepLab v3+: https://github.com/tensorflow/models/tree/master/research/deeplab

[6] UPerNet: https://github.com/CSAILVision/unifiedparsing

[7] HRNetV2: https://github.com/HRNet/HRNet-Semantic-Segmentation

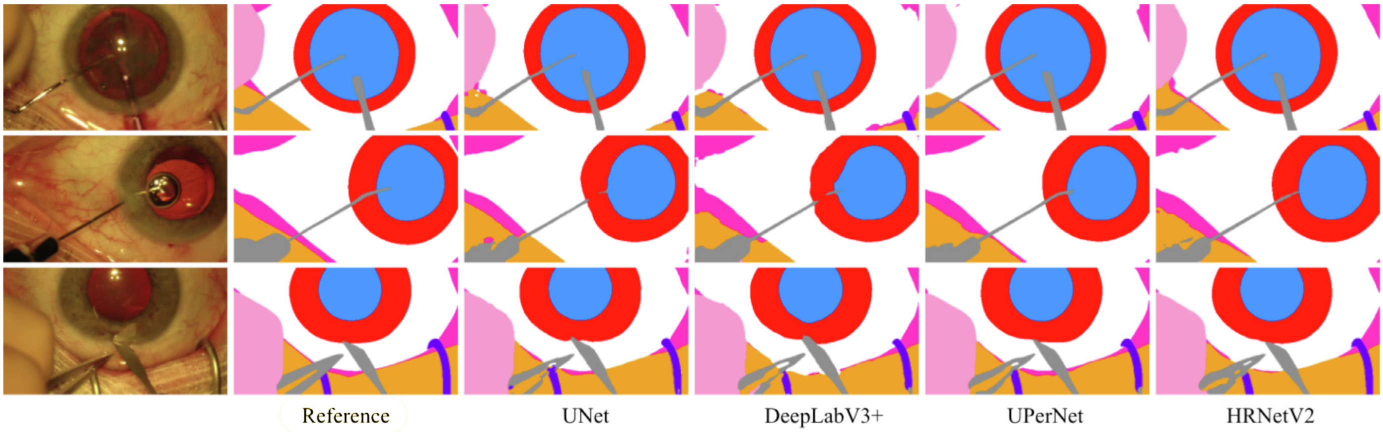


Fig. 3. Example frames with reference segmentation mask and model predictions for Task I. The top row presents a representative case for all models, while middle and bottom row depict two failure cases for all models. (Colormap: 🟦 Pupil, 🟥 Iris, ⬜ Cornea, 🟪 Skin, 🟧 Surgical tape, 🟪 Eye retractors and ⬛ Instrument).

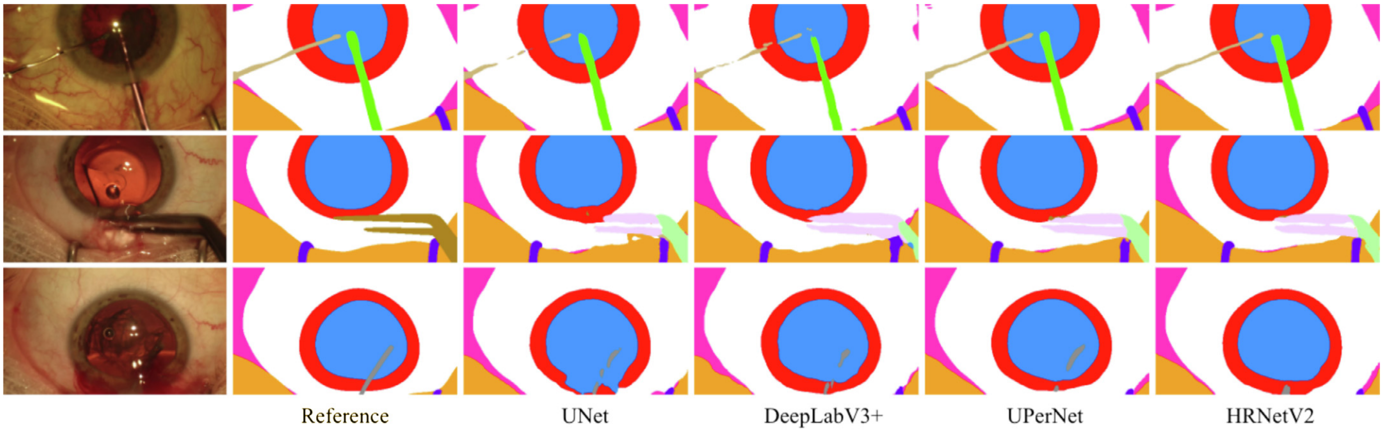


Fig. 4. Example frames with reference segmentation mask and model predictions for Task II. The top row presents a representative case for all models, while middle and bottom row depict two failure cases for all models. (Colormap: 🟦 Pupil, 🟥 Iris, ⬜ Cornea, 🟪 Skin, 🟧 Surgical tape, 🟪 Eye retractors, ⬛ Cannula, ⬜ Tissue forceps, 🟫 Capsulorhexis forceps, 🟩 Secondary knife, 🟫 Micromanipulator and 🟩 Phacoemulsification handpiece).

DeepLabV3+, UPerNet and HRNetV2 are 52%, 58.2%, 63% and 62.5% respectively (Table 9). For anatomy segmentation, the IoUs per class are similar to the previous tasks (Table 9).

3.7. Discussion

3.7.1. Task I

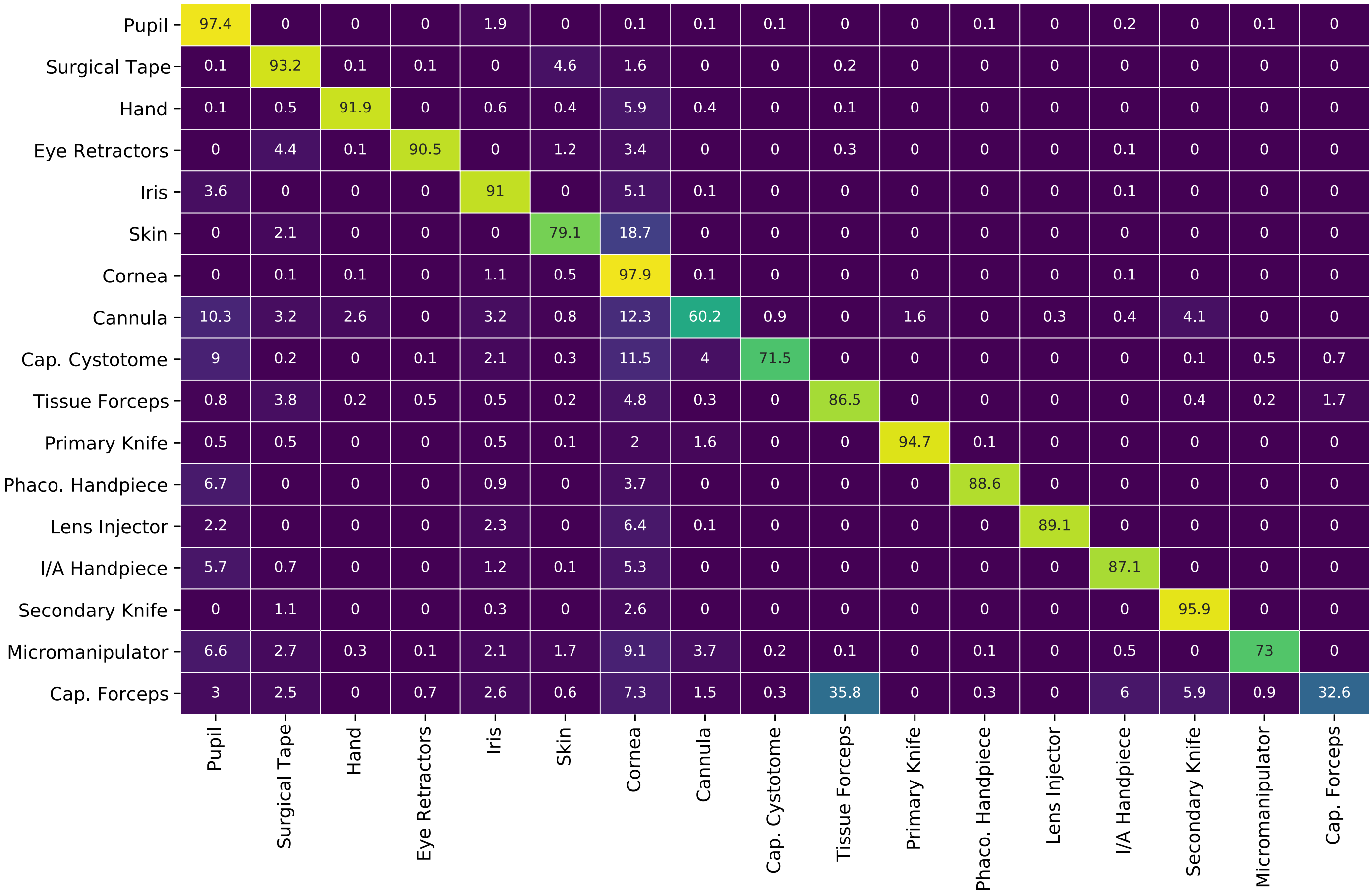
The mIoU for all models of Task I is comparable for the segmentation into 8 classes, with HRNetV2 presenting the highest mIoU and DeepLabV3+ the lowest for both validation and test sets (Table 4). The small differences in the mIoU between the models is because the imbalance among the classes is reduced by representing all instruments with one class. Therefore, since this experiment excluded instrument classification, it focused on scene segmentation. Fig. 3 (top row) shows a representative frame for this task, as in most frames all models performs well. Failure cases for this task include missed instruments tips as for the tip of the instrument in Fig. 3 (middle row) and inaccurate segmentation as for the handle of the instrument in Fig. 3 (middle row) and the forceps in Fig. 3 (bottom row).

3.7.2. Task II

The differences among the networks for simultaneous anatomy segmentation and multiple instrument segmentation are more visible as the number of classes increases. (Table 6). It can be seen that all networks achieve a high mIoU for large classes, such as the anatomical classes and instrument classes that are represented by large number of pixels (Table 7). For the instrument classes that appear in the test set, UNet has the lowest mIoU with 60.9% and HRNetV2 the highest with 68.56%. It is worth noting that finer instruments, such as the cannula and the micromanipulator have low IoUs (Table 7). This is because part of these instruments might be missed by the segmentation model, especially when they have been inserted in the anatomy. A representative frame that shows how finer instrument segmentation is more challenging for some of the models is depicted in Fig. 4 (top row). A failure cases for all models is shown in Fig. 4 (bottom row) where large part of the cannula has not been segmented by all models. These cases are also evident in the confusion matrix for HRNetV2, as it is shown that for example the cannula has also been segmented as either pupil or cornea (Fig. 6). Another failure case across models is spatially inconsistent instrument classification, where parts of the same instrument has been classified as different types, particularly between visually similar instruments (Fig. 4 - middle row). Similarly, this is also shown in the confusion matrix where the capsulorhexis forceps has been also classified as Bonn forceps.

3.7.3. Task III

The mIoU per class given in Table 9 shows that the instrument tips and handles are classified with varying degrees of accuracy. A representative case for all models is shown in Fig. 5 (top row) which shows how sensitive each model is to inconsistent



	Pupil	Surgical Tape	Hand	Eye Retractors	Iris	Skin	Cornea	Cannula	Cap. Cystotome	Tissue Forceps	Primary Knife	Phaco. Handpiece	Lens Injector	I/A Handpiece	Secondary Knife	Micromanipulator	Cap. Forceps
Pupil	97.4	0	0	0	1.9	0	0.1	0.1	0.1	0	0	0.1	0	0.2	0	0.1	0
Surgical Tape	0.1	93.2	0.1	0.1	0	4.6	1.6	0	0	0.2	0	0	0	0	0	0	0
Hand	0.1	0.5	91.9	0	0.6	0.4	5.9	0.4	0	0.1	0	0	0	0	0	0	0
Eye Retractors	0	4.4	0.1	90.5	0	1.2	3.4	0	0	0.3	0	0	0	0.1	0	0	0
Iris	3.6	0	0	0	91	0	5.1	0.1	0	0	0	0	0	0.1	0	0	0
Skin	0	2.1	0	0	0	79.1	18.7	0	0	0	0	0	0	0	0	0	0
Cornea	0	0.1	0.1	0	1.1	0.5	97.9	0.1	0	0	0	0	0	0.1	0	0	0
Cannula	10.3	3.2	2.6	0	3.2	0.8	12.3	60.2	0.9	0	1.6	0	0.3	0.4	4.1	0	0
Cap. Cystotome	9	0.2	0	0.1	2.1	0.3	11.5	4	71.5	0	0	0	0	0	0.1	0.5	0.7
Tissue Forceps	0.8	3.8	0.2	0.5	0.5	0.2	4.8	0.3	0	86.5	0	0	0	0	0.4	0.2	1.7
Primary Knife	0.5	0.5	0	0	0.5	0.1	2	1.6	0	0	94.7	0.1	0	0	0	0	0
Phaco. Handpiece	6.7	0	0	0	0.9	0	3.7	0	0	0	0	88.6	0	0	0	0	0
Lens Injector	2.2	0	0	0	2.3	0	6.4	0.1	0	0	0	0	89.1	0	0	0	0
I/A Handpiece	5.7	0.7	0	0	1.2	0.1	5.3	0	0	0	0	0	0	87.1	0	0	0
Secondary Knife	0	1.1	0	0	0.3	0	2.6	0	0	0	0	0	0	0	95.9	0	0
Micromanipulator	6.6	2.7	0.3	0.1	2.1	1.7	9.1	3.7	0.2	0.1	0	0.1	0	0.5	0	73	0
Cap. Forceps	3	2.5	0	0.7	2.6	0.6	7.3	1.5	0.3	35.8	0	0.3	0	6	5.9	0.9	32.6

Fig. 6. Confusion matrix for HRNetV2 on test set of Task II.

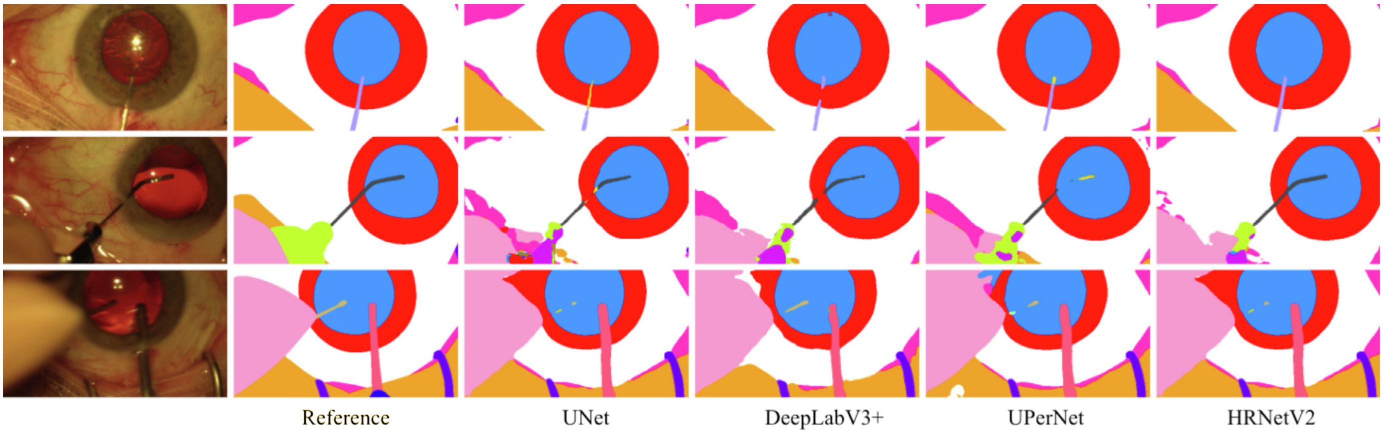


Fig. 5. Example frames with reference segmentation mask and model predictions for Task III. The top row presents a representative case for all models, while middle and bottom row depict two failure cases for all models. (Colormap: 🟦 Pupil, 🟥 Iris, ⬜ Cornea, 🟪 Skin, 🟧 Surgical tape, 🟪 Eye retractors, 🟨 Viscoelastic cannula, ⬛ Rycroft cannula, 🟩 Rycroft cannula handle, 🟪 Secondary knife handle, 🟪 Capsulorhexis cystotome, 🟫 Micromanipulator, 🟥 I/A handpiece and 🟦 I/A handpiece handle).

instrument classification and accurate segmentation. The main failure cases that are present in this task are spatially inconsistent instrument classification (Fig. 5 (middle row)) and missed parts of instruments (Fig. 5 (bottom row)). It is worth noting that these issues also appear in task II but to a lesser degree as task II segments the scene into fewer classes. In particular, in Fig. 5 (middle row), it can be seen that parts of the same instrument have been classified as multiple types of instruments (rycroft and viscoelastic cannula for the instrument tip and rycroft cannula handle and secondary knife handle for instrument handle in the middle row).

3.7.4. *All tasks*

In conclusion, the main challenges for the presented dataset across tasks and for all models are spatially consistent instrument classification and accurate segmentation of finer instruments, especially when instruments have been inserted into the anatomy.

There is a consistent difference between the mIoU for the validation and test sets as can be seen in Tables 4, 6 and 8. This can be explained by the distribution of class instances in each set, despite the attempt to have a similar distribution of instances at each set of videos, there is a variance in the distribution of instrument classes. This justifies further the choice of ignoring classes that do not have sufficient instances in the test set and are not present in the validation set.

Overall, DeepLabV3+, UPerNet and HRNetV2 achieve higher mIoU for instrument segmentation and classification than UNet (Tables 5, 7 and 9). In particular, UNet achieves a mIoU over 85% for anatomy segmentation in all tasks but gives a lower mIoU at instrument segmentation and classification. As mentioned, this difference in performance is smaller when the type of instrument does not need to be identified (Table 5) but is more evident when instrument classification is performed (Tables 7 and 9). UPerNet and HRNetV2 have the higher mIoU at simultaneous anatomy segmentation and instrument classification. It is also worth noting that DeepLabV3+ was trained using a MobileNetV2 backbone. This was to assess the performance of a light-weight version of the network. It performs more accurate instrument segmentation than UNet as the mIoU for instrument classes for all tasks highlights (Tables 5, 7 and 9).

4. Conclusions

Semantic segmentation of a surgical scene can improve understanding of the workflow of a surgical procedure and is crucial for intra-operative image guidance. In this paper, we present a dataset for semantic segmentation of images from cataract surgery procedures. The dataset consists of 4670 labelled images, which are sampled from the training set of the CATARACTS challenge dataset. The dataset labels include 36 classes and, in particular, four classes describing anatomical structures, 29 surgical instrument classes and three classes for other objects appearing in the surgical scene. The statistics presented for the dataset illustrate that the dataset is imbalanced, as the surgical instrument classes appear less frequently and are represented by fewer pixels compared to the anatomy classes. Three tasks were performed using the UNet, DeepLabV3+, UPerNet and HRNetV2 deep learning models. Each task presents different groups of instrument classes in order to assess the effect of simultaneous instrument classification on the segmentation output. It was shown that the four networks perform similarly for a relatively small number of classes with comparable number of pixels, addressing the imbalance issue. As the number of classes increase, HRNetV2 and UPerNet perform better in simultaneous anatomy segmentation and instrument classification than DeepLabV3+ and UNet, as HRNetV2 and UPerNet have a larger receptive field and are more capable of segmenting finer features. The mIoU per class metric reveals that UNet performs well in segmenting large areas such as the anatomical structures while DeepLabV3+, UPerNet and HRNetV2 provide more consistent instrument segmentation and classification in all performed tasks. Overall, the open challenges of the dataset are spatially consistent instrument classification, where parts of the same instrument can be classified as different types, and accurate segmentation of instruments, particularly when inserted into the anatomy. The aim of introducing a dataset for semantic segmentation in cataract surgery is to facilitate further development of computer-assisted strategies for image guidance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Maria Grammatikopoulou: Methodology, Software, Data curation, Writing - original draft. **Evangello Flouty:** Data curation, Writing - original draft. **Abdolrahim Kadkhodamohammadi:** Software, Writing - review & editing. **Gwenolé Quellec:** Resources. **Andre Chow:** Funding acquisition. **Jean Nehme:** Funding acquisition. **Imanol Luengo:** Software, Supervision, Writing - review & editing. **Danail Stoyanov:** Funding acquisition, Project administration, Writing - review & editing.

Acknowledgements

We would like to thank Ellie Jaram, Nunzia Lombardo, Fanni Demeter and Hannah Bradd for their efforts in annotating the dataset to the highest quality.

This work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) at UCL (203145Z/16/Z), EPSRC (EP/P012841/1, EP/P027938/1, EP/R004080/1) and the H2020 FET (GA 863146). Danail Stoyanov is supported by a Royal Academy of Engineering Chair in Emerging Technologies (CiET18196) and an EPSRC Early Career Research Fellowship (EP/P012841/1).

References

Al Hajj, H., et al., 2019. Cataracts: challenge on automatic tool annotation for cataract surgery. Med. Image Anal. 52, 24–41.
Allan, M., et al., 2015. Image based surgical instrument pose estimation with multi--class labelling and optical flow. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 331–338.
Allan, M., et al., 2019. 2017 Robotic Instrument Segmentation Challenge. arXiv preprint arXiv:1902.06426.
Allan, M., et al., 2020. 2018 Robotic Scene Segmentation Challenge. arXiv preprint arXiv:2001.11190.
Bouget, D., et al., 2019. Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in ct data for lung cancer staging. Int. J. Comput. Assist. Radiol. Surg. 14 (6), 1–10.
Chakrabarti, A., Nazm, N., 2017. Posterior capsular rent: prevention and management. Indian J. Ophthalmol. 65 (12), 1359.
Chen, L.-C., et al., 2017. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.
Chen, L.-C., et al., 2017b. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.
Chen, L.-C., et al., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818.
Chollet, F., 2017. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258.
Deng, J., et al., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.
Engelhardt, S., et al., 2018. Improving surgical training phantoms by hyperrealism: deep unpaired image-to-image translation from real surgeries. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 747–755.
Everingham, M., et al., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88 (2), 303–338.
Fuentes-Hurtado, F., et al., 2019. Easylabels: weak labels for scene segmentation in laparoscopic videos. Int. J. Comput. Assist. Radiol. Surg. 14 (7), 1–11.
García-Peraza-Herrera, L.C., et al., 2016. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In: International Workshop on Computer-Assisted and Robotic Endoscopy. Springer, pp. 84–95.
Kassahun, Y., et al., 2016. Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. Int. J. Comput. Assist. Radiol. Surg. 11 (4), 553–568.
Kim, J.E., et al., 2012. Medical malpractice claims related to cataract surgery complicated by retained lens fragments (an American ophthalmological society thesis). Trans. Am. Ophthalmol. Soc. 110, 94.
Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kovler, I., et al., 2015. Haptic computer-assisted patient-specific preoperative planning for orthopedic fractures surgery. Int. J. Comput. Assist. Radiol. Surg. 10 (10), 1535–1546.
Maier-Hein, L., et al., 2017. Surgical data science: enabling next-generation surgery. Nat. Biomed. Eng. 1, 691–696.
Maier-Hein, L., et al., 2020. Bias: Transparent reporting of biomedical image analysis challenges. Med. Image Anal. 66, 101796.
Ozdemir, F., Goksel, O., 2019. Extending pretrained segmentation networks with additional anatomical structures. Int. J. Comput. Assist. Radiol. Surg. 14 (7), 1–9.

Pfeiffer, M., et al., 2019. Learning soft tissue behavior of organs for surgical navigation with convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. 14 (7), 1147–1155.

Ronneberger, O., et al., 2015. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Sandler, M., et al., 2018. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520.

Suetens, P., et al., 1993. Image segmentation: methods and applications in diagnostic radiology and nuclear medicine. Eur. J. Radiol. 17 (1), 14–21.

Sznitman, R., et al., 2012. Data-driven visual tracking in retinal microsurgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 568–575.

Twinanda, A.P., et al., 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging 36 (1), 86–97.

Vedula, S.S., et al., 2017. Objective assessment of surgical technical skill and competency in the operating room. Annu. Rev. Biomed. Eng. 19, 301–325.

Wang, J., et al., 2020. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell..

WHO, 2018. Priority eye diseases. https://www.who.int/blindness/causes/priority/en/index1.html.

Xiao, T., et al., 2018. Unified perceptual parsing for scene understanding. In: European Conference on Computer Vision. Springer.

Zeng, C., et al., 2016. A combination of three-dimensional printing and computer-assisted virtual surgical procedure for preoperative planning of acetabular fracture reduction. Injury 47 (10), 2223–2227.

Zisimopoulos, O., et al., 2018. Deepphase: surgical phase recognition in cataracts videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 265–272.