Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 226 (2023) 36–42

Procedia Computer Science

www.elsevier.com/locate/procedia

Proceedings of International Conference on Biomimetic, Intelligence and Robots

Landmark Detection using Transformer Toward Robot-assisted Nasal Airway Intubation

Tianhang Liu[a], Hechen Li[a], Long Bai[a], Yanan Wu[a,b,c], An Wang[a], Mobarakol Islam[d], Hongliang Ren[a,e,f,g,*]

[a]*Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China*
[b]*College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China*
[c]*Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, Shenyang, China*
[d]*Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, London, UK*
[e]*Department of Biomedical Engineering, National University of Singapore, Singapore*
[f]*Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong, Hong Kong, China*
[g]*Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China*

Abstract

Robot-assisted airway intubation application needs high accuracy in locating targets and organs. Two vital landmarks, nostrils and glottis, can be detected during the intubation to accommodate the stages of nasal intubation. Automated landmark detection can provide accurate localization and quantitative evaluation. The Detection Transformer (DeTR) leads object detectors to a new paradigm with long-range dependence. However, current DeTR requires long iterations to converge, and does not perform well in detecting small objects. This paper proposes a transformer-based landmark detection solution with deformable DeTR and the semantic-aligned-matching module for detecting landmarks in robot-assisted intubation. The semantics aligner can effectively align the semantics of object queries and image features in the same embedding space using the most discriminative features. To evaluate the performance of our solution, we utilize a publicly accessible glottis dataset and automatically annotate a nostril detection dataset. The experimental results demonstrate our competitive performance in detection accuracy. Our code can be accessible at https://github.com/ConorLTH/airway_intubation_landmarks_detection.

Keywords: Bleeding regions segmentation; Medical image segmentation; Semi-supervised learning; Video capsule endoscopy

* Corresponding author
E-mail address: hlren@ee.cuhk.edu.hk

1877-0509
10.1016/j.procs.2023.10.633

1. Introduction

Object detection is critical in computer vision, necessitating accurate recognition and localization of objects within frames. Deep learning methods have spurred the creation of numerous models that can detect objects efficiently and effectively. Recently, Detection Transformer (DeTR) [5] was introduced, leading to novel design methodologies. Before DeTR, human-designed elements and detection processes heavily impacted detector performance. However, DeTR enables fully end-to-end detection frameworks. Nevertheless, DeTR's performance is limited by two issues: low accuracy for detecting small objects and slow convergence rates.

To address these limitations, researchers developed Deformable DeTR [30]. Inspired by Deformable Convolution [11, 30], this model provides better performance on small objects. Researchers noted that DeTR's slow convergence was primarily due to challenges in matching object queries with features, particularly during cross-attention in the decoder. To address this challenge, they suggested methods that would help object queries match with objects' features. For instance, Conditional DeTR [20] splits the object query into query content and query spatial embedding to search relevant areas based on objects' appearances and locate extremity areas. Spatially modulated co-attention (SMCA) DeTR [12] introduces a Gaussian spatial map that forces cross-attention to focus on specific local spatial areas. Semantic-Aligned-Matching (SAM) DeTR [26] designs a flexible module that implements semantics-align to simplify the matching of objects' features and object queries.

The development of computer vision has propelled its application in many medical fields, such as image preprocessing [2, 9, 10], medical diagnosis [4, 6, 7], robot-assisted surgery and intubation [3, 24, 27]. One such application is airway intubation, which may require nasal intubation for semiconscious or awake patients when a laryngoscope is not an option but poses challenges due to insufficient space. Intubation during medical procedures can pose significant challenges. To overcome these challenges, some researchers have implemented convolutional neural networks (CNNs) to classify and detect intubation difficulties [25, 28]. For example, Hayasaka *et al.* [15] developed an AI model that uses the patient's facial image to classify intubation difficulty. Similarly, Aguilar *et al.* [1] proposed a mobile app that employs a CNN to detect a difficult airway. Specifically, this work defines two classes for a challenging airway based on the Mallampati score. Deep learning in difficult airway detection has shown promise in facilitating intubation procedures. However, there are several areas for improvement. For instance, existing models focus solely on detecting difficult airways, without additional information on intubation procedures, which can limit their practicality. Therefore, a more comprehensive approach to difficult airway detection is necessary. However, the need for medical professionals to label these datasets manually has limited further progress. Moreover, creating datasets for specific applications based on current open-source datasets and automatically generating annotations can overcome this limitation.

To overcome these limitations, we present a novel approach for achieving comprehensive landmark detection to enhance the efficiency of intubation procedures. Landmarks serve as essential indicators during the intubation process, providing valuable information about its progress. The specific landmarks to be detected may vary at different stages of the intubation process. For example, during the initial phase, when the tube is external to the human body, detecting the position of the nostril becomes essential. Subsequently, as the tube moves inside the nasal passage, detecting the glottis becomes crucial. However, some landmarks, such as nostrils, are relatively small objects, presenting a challenge for accurate detection by methods like DeTR. In this work, we make the following contributions to address these challenges:

- We propose a comprehensive solution to facilitate the successful execution of robot-assisted intubation. Our method demonstrates the capability to effectively detect pertinent anatomical landmarks, thereby furnishing critical information required for precise intubation procedures.
- We integrate the semantic aligner and deformable attention module, into the detection transformer framework for landmark detection tasks. This integration enhances the model's ability to identify and localize landmarks throughout the intubation process accurately.
- Additionally, we devise an automatic annotation methodology for generating detection bounding boxes for a nostril dataset. Furthermore, through extensive experimentation, we provide compelling evidence showcasing the superior performance of our proposed solution in accurately detecting landmarks during intubation. Our transformer-based detection solution achieves outstanding performance with only 24-epoch training.

2. Detection Model

In this section, we will discuss the DeTR model, its current challenges, and the process of adapting DeTR for the practical application of airway landmark detection.

2.1. DeTR and its Challenges

DeTR is a state-of-the-art object detection model that treats object detection as a set prediction problem. It uses learnable object queries to form relations with extracted features using the transformer-based encoder and decoder modules. In the encoder, self-attention modules encode the extracted image features. In the decoder, two attention mechanisms - self-attention modules and cross-attention modules - are utilized to enable information exchange between object queries and encoded features. DeTR faces two challenges. Firstly, the computation complexity of the attention mechanism increases with the spatial size of the input, making it difficult to implement multi-level feature maps and improve small object detection accuracy [30]. Secondly, the random initialization of object queries makes them pair equally with all spatial locations of the image, rather than specific regions for detecting objects [13, 20, 30]. This prevents training object queries from focusing on specific regions. Therefore, long training iterations are necessary for DeTR to perform well.

To address the computation complexity challenge, Deformable DeTR was proposed. It combines deformable convolution with the attention mechanism to create Multi-scale Deformable Attention [30]. This novel approach can aggregate features from multi-level feature maps and decrease the computation complexity of the original attention mechanism used in DeTR. The deformable attention only samples key positions on the feature map to act as queries, rather than considering all possible spatial locations.

Another variant, SAM-DeTR, proposes the Semantic Aligner Module to accelerate DeTR's convergence. To ensure semantic alignment, it aligns object queries and features within uniform embedding spaces. SAM uses learnable reference boxes to guide object queries to align with the semantics of encoded features. This process involves three sub-operations: extracting region features with reference boxes using ROIAlign; resampling salient points with M points for each region to form new object queries and their corresponding positional embeddings; and evaluating reweighting coefficients with current object queries to combine messages of old and new object queries. The operations related to object queries are formulated using the sigmoid function.

2.2. Landmark Detection Solution

We propose integrating the semantic aligner module into Deformable DeTR, which combines deformable convolution with multi-scale deformable attention. To achieve this integration, we focus on adapting the decoder in Deformable DeTR to accommodate the SAM module. Specifically, we model the object queries and their positional embeddings as $query$, $query_{pos} \in \mathbb{R}^{N,256}$, following Deformable DeTR's implementation, instead of directly modeling the object query of $\mathbb{R}^{N,4}$ used in the SAM-DeTR for modeling reference boxes where N represents the number of the object queries and 4 and 256 are hyperparameters for the channel of queries. We also use four feature levels in multi-scale deformable attention modules, consistent with the implementation in Deformable DeTR. In contrast, the SAM module of each layer in the decoder only uses encoded features from a single level to realize semantic alignment between object queries and encoded features. To achieve this, different decoder layers take encoded features from different levels as input.

The reference boxes used in the SAM module are obtained via a learnable linear projection over query position embeddings, while the reference points used in the cross-attention modules of the decoder are obtained from a learnable linear projection over reference boxes. To avoid inappropriate reference points being chosen, we do not directly regard the center points as the reference points, even though the reference boxes are represented as coordinates of the center points and the height and width of the bounding boxes.

To ensure interoperability between the cross-attention and SAM modules, we set the hidden dimension of cross-attention to 256×8. Here, 256 is the dimension of the object queries, and 8 is the number of heads in multi-head attention. In the SAM module, each object query is assigned a reference box. During resampling, each reference box is resampled with M salient points. M representing the number of heads in multi-head attention (typically 8). Thus, after resampling, the dimension of the query changes to $Q^{new} \in \mathbb{R}^{N \times M \times 256}$. Before sending Q^{new} into the cross-attention

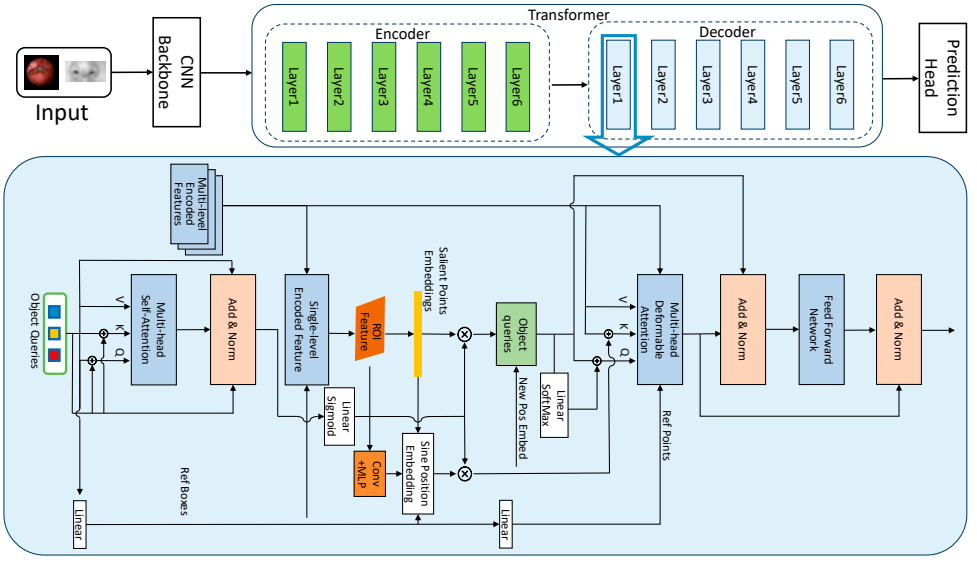


Fig. 1. The workflow of our proposed solution with deformable attention and semantic aligner.

module, it is reshaped to $[N, M \times 256]$. Therefore, our approach integrates the SAM module into Deformable DeTR, providing a more effective and efficient landmark detection model. The overview architecture of our solution can be found in Fig. 1.

3. Experiments

3.1. Datasets

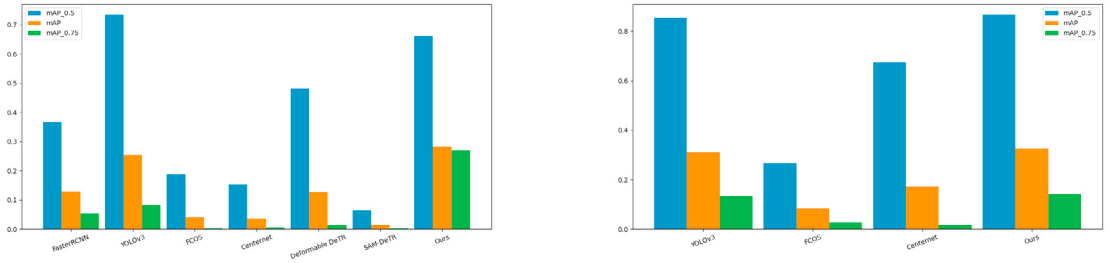
In the context of this research, we utilized two distinct datasets: one for nostril detection and another for glottis detection, with the aim of enabling airway intubation landmark detection.

The nostril dataset is derived from the publicly available BioID dataset [17], which was initially collected and annotated for human face keypoint detection. The dataset consists of 1521 grayscale images, all standardized to a resolution of 384x286 pixels. We manually use key point annotations of nostrils provided by the BioID dataset as the center points and expand the height and width to form bounding box annotations. To denote the nostril locations, we adopt the bounding box annotation format following the COCO format [19].

Similarly, the glottis dataset is based on the publicly available BAGALS dataset [14], initially designed for glottis segmentation. Collaboratively compiled by seven institutions, the BAGALS dataset comprises endoscope videos, detailed segmentation annotations, and frame-level information. From this collection, we extracted a subset of explicit annotations from nasal endoscopic videos, resulting in 881 images for glottis detection. As the videos originate from diverse sources, the frames within the subset exhibit varying resolutions. To form corresponding bounding box annotations, we use OpenCV to capture the contours of the segmentation annotations of BAGALS and annotate with their maximum and minimum coordinate values. We follow the COCO format for annotating the glottis dataset to maintain consistency.

3.2. Implementation Details

In our experiments, the object detection models are implemented using MMdetection [8], a freely available toolbox for PyTorch that supports various deep learning-based detectors. All models are trained from scratch and evaluated using the abovementioned two datasets. The nostril dataset is partitioned into 1021 frames for training, 185 frames for validation, and 315 frames for testing. Similarly, the glottis dataset is divided into 377 frames for training, 88 frames



(a) mAP on glottis dataset (b) mAP on nostril dataset

Fig. 2. The visualization of the mAP results from the comparison experiments on the glottis and nostril dataset.

for validation, and 416 frames for testing. For feature extraction, we opt for the ResNet50 backbone [16]. Besides, we employed the Adam optimizer [18] with a learning rate of 1×10^{-5} for the backbone and 1×10^{-4} for the detectors. We ensure standardized implementation and evaluation across all models by utilizing MMdetection and adhering to this experimental setup, facilitating reliable comparisons and analysis. We set the number of epochs to 24 to observe the performance under limited training iterations.

3.3. Experimental Results

Table 1. Comparison experiments of our SAM Deformable DeTR against baseline models on the glottis dataset.

Method	mAP	mAP@0.5	mAP@0.75
FasterRCNN [22]	0.129	0.366	0.054
YOLOv3 [21]	0.254	**0.733**	0.084
FCOS [23]	0.042	0.188	0.002
Centernet [29]	0.037	0.153	0.006
Deformable DeTR [30]	0.128	0.482	0.013
SAM-DeTR [26]	0.015	0.064	0.003
Ours	**0.282**	0.661	**0.270**

Table 2. Comparison experiments of our SAM Deformable DeTR against baseline models on the nostril dataset.

Method	mAP	mAP@0.5	mAP@0.75
YOLOv3 [21]	0.311	0.855	0.133
FCOS [23]	0.084	0.267	0.027
Centernet [29]	0.171	0.673	0.017
Ours	**0.325**	**0.865**	**0.142**

Tables 1 and 2 present the results obtained from various models trained on two distinct datasets using a consistent 24-epoch training scheme. The baseline models consist of both CNN-based and Transformer-based architectures. The performance evaluation adheres to the standard COCO metric. We also visualize the qualitative results in Fig. 3.

As depicted in Table 1 and Figure 2(a), the test results on the glottis dataset reveal the advantages of combining the semantics aligner module with deformable attention, as evident from the performance of SAM Deformable DeTR and other baseline models trained from scratch. Notably, SAM-DeTR exhibits inferior results on the glottis dataset. This can be attributed to the inherent limitations of DeTR, despite the inclusion of a plug-and-play module in SAM-DeTR to expedite its convergence. SAM-DeTR encounters similar challenges as DeTR when trained on datasets with limited capacity. In contrast, the performance of Deformable DeTR surpasses that of SAM-DeTR, although it falls short compared to certain CNN-based detectors. These findings suggest that integrating the semantic aligner module

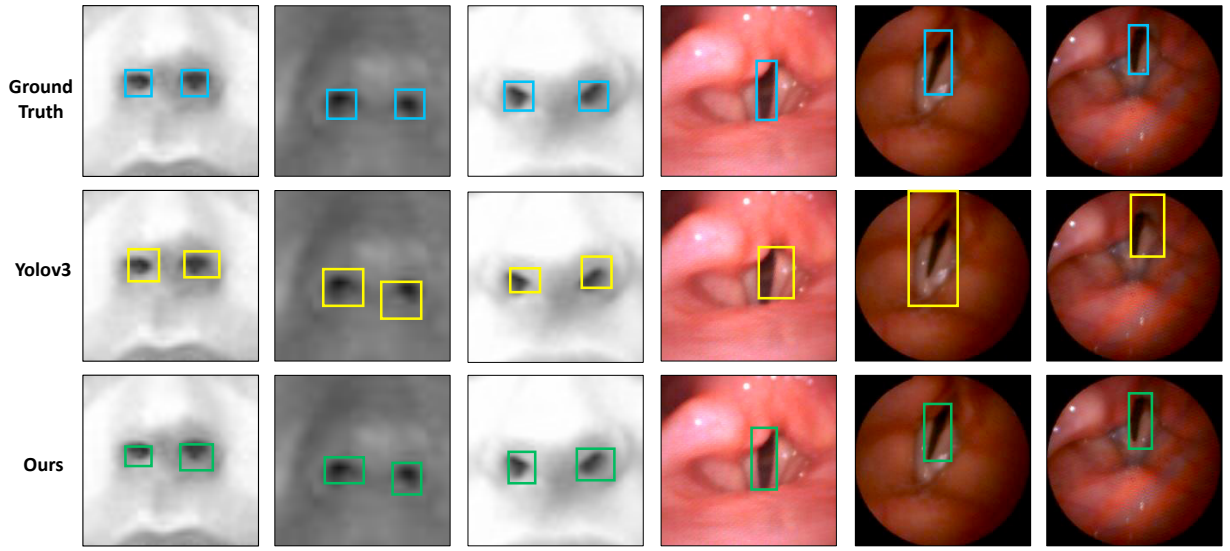


Fig. 3. The qualitative comparison of the experimental results on the nostril (column 1-3) and glottis (column 4-6) detection task.

into Deformable DeTR has the potential to enhance the model's capabilities, an observation supported by the results of SAM Deformable DeTR. Specifically, the mAP increment demonstrates an 18-fold increase compared to SAM-DeTR and a 2-fold increase compared to Deformable DeTR.

In Table 2 and Figure 2(b), the performance of SAM Deformable DeTR on the nostril dataset achieves the highest capability. The Deformable DeTR and SAM-DeTR results are omitted from Table 2 since their mAP results are 0.000. This suggests that these models were not fully trained on the nostril dataset with limited training iterations. The nostril dataset lacks distinct appearances, and the features of the nostril are strongly correlated with the fixed locations and surrounding facial organs. SAM Deformable DeTR effectively leverages multi-scale features and spatial information in object detection, contributing to its superior performance on this dataset.

4. Conclusion

In this research work, we employ SAM Deformable DeTR, an improved version of the Deformable DeTR framework integrated with the semantics aligner module inspired by SAM-DeTR. Our primary objective is to evaluate the performance of our solution and demonstrate its effectiveness in detecting airway intubation landmarks for medical applications. Comprehensive experiments were conducted using two public datasets, encompassing various scenarios encountered in airway intubation procedures. The experimental results indicate that SAM Deformable DeTR surpasses both the original SAM-DeTR framework and the standard Deformable DeTR model regarding detection accuracy. The outcomes of our study demonstrate the significant advancements of our solution in the domain of airway intubation landmark detection. The proposed framework exhibits superior performance compared to existing methods, showcasing its potential for improving medical interventions and facilitating accurate and efficient airway-related procedures.

Acknowledgements

This work was supported by Hong Kong Research Grants Council (RGC) Research Impact Fund (RIF) R4020-22, Collaborative Research Fund (CRF C4026-21GF, CRF C4063-18G), General Research Fund (GRF 14203323, GRF 14216022, and GRF 14211420), NSFC/RGC Joint Research Scheme N_CUHK420/22, GRS #3110167; Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) STIC Grant SGDX20210823103535014 (202108233000303); Guangdong Basic and Applied Basic Research Foundation (GBABF) #2021B1515120035; Shun Hing Institute of Advanced Engineering (SHIAE Project BME-p1-21) at The Chinese University of Hong Kong.

References

[1] Aguilar, K., Alférez, G.H., Aguilar, C., 2020. Detection of difficult airway using deep learning. Machine Vision and Applications 31, 1–11.

[2] Bai, L., Chen, T., Wu, Y., Wang, A., Islam, M., Ren, H., 2023a. Llcaps: Learning to illuminate low-light capsule endoscopy with curved wavelet attention and reverse diffusion. arXiv preprint arXiv:2307.02452 .

[3] Bai, L., Islam, M., Seenivasan, L., Ren, H., 2023b. Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. arXiv preprint arXiv:2305.11692 .

[4] Bai, L., Wang, L., Chen, T., Zhao, Y., Ren, H., 2022. Transformer-based disease identification for small-scale imbalanced capsule endoscopy dataset. Electronics 11, 2747.

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer. pp. 213–229.

[6] Che, H., Chen, S., Chen, H., 2023. Image quality-aware diagnosis via meta-knowledge co-embedding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19819–19829.

[7] Che, H., Jin, H., Chen, H., 2022. Learning robust representation for joint grading of ophthalmic diseases via adaptive curriculum and feature disentanglement, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III, Springer. pp. 523–533.

[8] Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al., 2019. Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 .

[9] Chen, Z., Gu, S., Lu, G., Xu, D., 2022a. Exploiting intra-slice and inter-slice redundancy for learning-based lossless volumetric image compression. IEEE Transactions on Image Processing 31, 1697–1707.

[10] Chen, Z., Lu, G., Hu, Z., Liu, S., Jiang, W., Xu, D., 2022b. Lsvc: a learning-based stereo video compression framework, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6073–6082.

[11] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks, in: Proceedings of the IEEE international conference on computer vision, pp. 764–773.

[12] Gao, P., Zheng, M., Wang, X., Dai, J., Li, H., 2021a. Fast convergence of detr with spatially modulated co-attention, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3621–3630.

[13] Gao, P., Zheng, M., Wang, X., Dai, J., Li, H., 2021b. Fast convergence of detr with spatially modulated co-attention, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3621–3630.

[14] Gómez, P., Kist, A.M., Schlegel, P., Berry, D.A., Chhetri, D.K., Dürr, S., Echternach, M., Johnson, A.M., Kniesburges, S., Kunduk, M., et al., 2020. Bagls, a multihospital benchmark for automatic glottis segmentation. Scientific data 7, 186.

[15] Hayasaka, T., Kawano, K., Kurihara, K., Suzuki, H., Nakane, M., Kawamae, K., 2021. Creation of an artificial intelligence model for intubation difficulty classification by deep learning (convolutional neural network) using face images: an observational study. Journal of Intensive Care 9, 1–14.

[16] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition , 770–778.

[17] Jesorsky, O., Kirchberg, K.J., Frischholz, R.W., 2001. Robust face detection using the hausdorff distance, in: Audio-and Video-Based Biometric Person Authentication: Third International Conference, AVBPA 2001 Halmstad, Sweden, June 6–8, 2001 Proceedings 3, Springer. pp. 90–95.

[18] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

[19] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer. pp. 740–755.

[20] Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J., 2021. Conditional detr for fast training convergence, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3651–3660.

[21] Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 .

[22] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28.

[23] Tian, Z., Shen, C., Chen, H., He, T., 2019. Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9627–9636.

[24] Wang, G., Ren, T.A., Lai, J., Bai, L., Ren, H., 2023. Domain adaptive sim-to-real segmentation of oropharyngeal organs. arXiv preprint arXiv:2305.10883 .

[25] Wu, Y., Zhao, S., Qi, S., Feng, J., Pang, H., Chang, R., Bai, L., Li, M., Xia, S., Qian, W., et al., 2022. Two-stage contextual transformer-based convolutional neural network for airway extraction from ct images. arXiv preprint arXiv:2212.07651 .

[26] Zhang, G., Luo, Z., Yu, Y., Cui, K., Lu, S., 2022a. Accelerating detr convergence via semantic-aligned matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 949–958.

[27] Zhang, Y., Bai, L., Liu, L., Ren, H., Meng, M.Q.H., 2022b. Deep reinforcement learning-based control for stomach coverage scanning of wireless capsule endoscopy, in: 2022 IEEE International Conference on Robotics and Biomimetics (ROBIO), IEEE. pp. 01–06.

[28] Zhao, S., Wu, Y., Tong, M., Yao, Y., Qian, W., Qi, S., 2022. Cot-xnet: contextual transformer with xception network for diabetic retinopathy grading. Physics in Medicine & Biology 67, 245003.

[29] Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as points. arXiv preprint arXiv:1904.07850 .

[30] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2021. Deformable detr: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations (ICLR).