F2PASeg: Feature Fusion for Pituitary Anatomy Segmentation in Endoscopic Surgery

Lumin Chen[1],Zhiying Wu[1*], Tianye Lei[2], Xuexue Bai[3], Ming Feng[3], Yuxi Wang[1], Gaofeng Meng[1], Zhen Lei[1], and Hongbin Liu[1]

[1]Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science Innovation, Chinese Academy of Sciences
zhiying.wu@cair-cas.org.hk
[2]The University of Hong Kong
[3]Peking Union Medical College Hospital, Chinese Academy of Medical Science

Abstract. Pituitary tumors often cause deformation or encapsulation of adjacent vital structures. Anatomical structure segmentation can provide surgeons with early warnings of regions that pose surgical risks, thereby enhancing the safety of pituitary surgery. However, pixel-level annotated video stream datasets for pituitary surgeries are extremely rare. To address this challenge, we introduce a new dataset for Pituitary Anatomy Segmentation (PAS). PAS comprises 7,845 time-coherent images extracted from 120 videos. To mitigate class imbalance, we apply data augmentation techniques that simulate the presence of surgical instruments in the training data. One major challenge in pituitary anatomy segmentation is the inconsistency in feature representation due to occlusions, camera motion, and surgical bleeding. By incorporating a Feature Fusion module, F2PASeg is proposed to refine anatomical structure segmentation by leveraging both high-resolution image features and deep semantic embeddings, enhancing robustness against intraoperative variations. Experimental results demonstrate that F2PASeg consistently segments critical anatomical structures in real time, providing a reliable solution for intraoperative pituitary surgery planning. Code: https://github.com/paulili08/F2PASeg.

Keywords: Pitutary anatomy segmentation · Segment Anything Model · Surgical Vision.

1 Introduction

Automatic segmentation of anatomical structures can identify dangerous areas, surgical risks can be reduced in pituitary surgery [15, 20, 21]. Especially during the sella phase, anatomical structure segmentation is crucial due to the close proximity of the anatomy [15]. As dangerous areas are difficult to segment, surgeons face challenges when performing endoscopic pituitary surgery. The sella, where the pituitary tumor is located, can be accessed safely. However, the presence of internal carotid arteries and optic nerves beneath the smaller surrounding structures complicates the process of opening. The pituitary tumor leads to

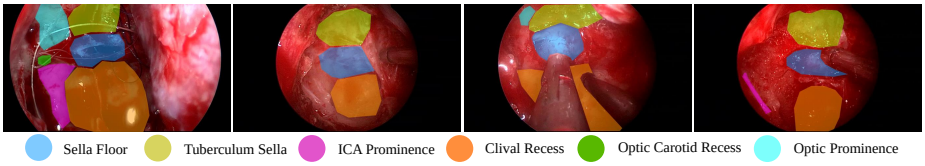

Fig. 1. Examples of the proposed PAS dataset including images and labels.

Table 1. Comparison with the existing datasets in pituitary anatomy segmentation.

Pituitary Anatomy Datasets	Image	Case	Class	Task
Sarwin et al. [18]	19000	166	16	Detection
Staartjes et al. [19]	549	23	3	Detection
Adrito et al. [8]	635	64	10	Segmentaion
Our PAS	7845	120	6	Segmentaion

compression, distortion, or encasement of the surrounding structures [2]. Inaccurate segmentation of essential anatomical structures can cause injury to the patients [14, 15].

The availability of anatomical datasets for pituitary surgeries is extremely limited. The scarcity of comprehensive and diverse datasets in pituitary surgeries poses a significant challenge during the sellar phase. Due to the considerable variations in anatomical structures among patients, collecting a large-scale dataset is crucial. Fig. 1 illustrates the semantic segmentation of six anatomical structures during the sellar phase of endoscopic pituitary surgery. As shown in Table 1, our dataset consists of 7,845 images. Compared to [18] and [19], bounding boxes/centroids annotated datasets produced for target detection task, our dataset provides pixel-level mask annotations for semantic segmentation, which is much more labor-intensive. Specifically, in the sellar phase, the number of images in our dataset is 12 times greater than that of the dataset in [8]. Moreover, our dataset includes nearly twice as many cases as the dataset presented in [8], with images in each case exhibiting a high degree of continuity, making them suitable for video-based analysis. Additionally, our dataset captures the significant variations in anatomical structures among patients and provides a comprehensive representation of the diverse anatomical scenarios encountered in pituitary surgeries.

A number of commonly used deep learning methods have been used for intraoperative endoscopic segmentation [9, 21, 22]. U-Net [17] uses weakly supervised learning of centroids to generate segmentation masks for each structure. In [8], U-Net++ is employed for the semantic segmentation of the two most prominent, largest, and frequently occurring structures (sella and clival recess) and for centroid detection of the remaining eight less frequently occurring structures. Recently, the field of semantic segmentation has undergone a significant shift with the increasing focus on large-scale pre-trained models. Segment Anything Model (SAM) [10], a leading Vision Transformer-based segmentation framework, has

made remarkable progress in expanding the boundaries of segmentation in natural images. The subsequent update of Segment Anything Model 2 (SAM2) [16] enables efficient video segmentation by transferring prompts with frame-to-frame continuity. It has largely achieved the end-to-end efficient segmentation required for intraoperative endoscopy, the current segmentation methods have not fully investigated feature fusion. Thus, we purpose F2PASeg, a video-based segmentation model enhancing feature fusion for complex scenes. The main contributions are summarized as follows:

1. We introduce the Pituitary Anatomy Segmentation (PAS) dataset, a large-scale collection consisting of 7,845 pixel-level annotated images captured during the sellar phase of endoscopic pituitary surgery. Our dataset PAS contains the significant variations in anatomical structures among patients.
2. We propose an efficient architecture F2PASeg for pituitary anatomy segmentation in endoscopic surgery. In our F2PASeg, a feature fusion module enhances the mask decoder's abilities to integrate image embeddings with high-dimensional features from the image encoder to optimize feature integration.
3. To address imbalanced distributions of critical structures, we multiplex surgical instrument annotations from the same dataset for data augmentation. In particular, for the sparsely distributed categories of pituitary anatomy, the original image frames are augmented with simulated surgical scenes that involve the use of surgical instruments.

2 Methods

2.1 Overview

In this section, we build an end-to-end promptable structure F2PASeg and augment the dataset for pituitary tumor surgery scenarios. First, we integrate a feature fusion module in mask decoder. This module combines two residual blocks with LoRa branch, which enhances the combinations between the features from image encoder and the embedding from memory encoder. With LoRA, the model better satisfies intraoperative real-time segmentation demands, achieving higher FPS and reduced parameters. Second, we augment for samples with small distributions by multiplexing the dataset to target less distributed samples, mitigating the effect of imbalanced distributions.

2.2 F2PASeg backbone

The Segment Anything Model (SAM) [10] has demonstrated strong performance in image segmentation tasks. However, its heavy reliance on prompts makes it unsuitable for intraoperative endoscopic surgery scenarios. The later version, SAM2 [16], extends image segmentation to the video domain and generates mask predictions across an entire video by leveraging a newly developed memory mechanism. As illustrated in Fig. 2(a), the proposed F2PASeg incorporates three

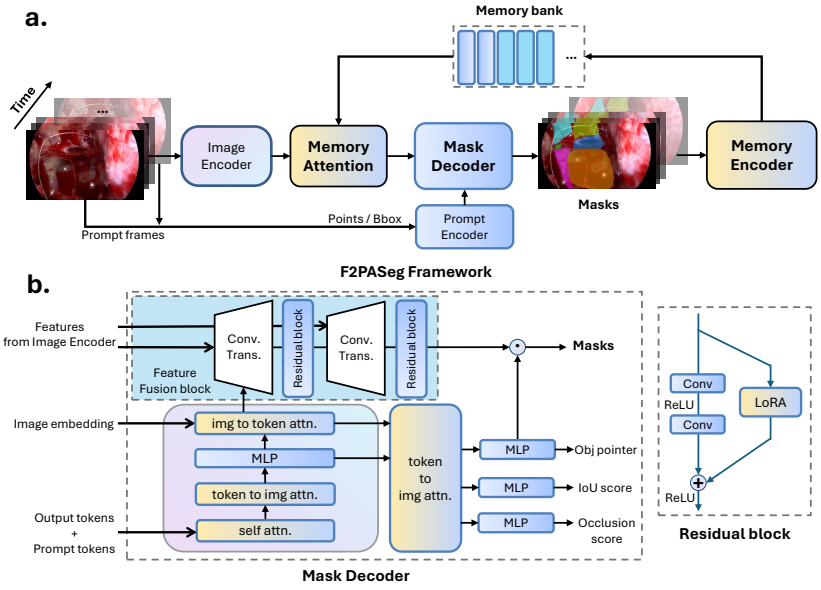

Fig. 2. The structure of our proposed F2PASeg model. (a) is the overall framework that contains image encoder, prompt encoder, memory attention, mask decoder and memory encoder. (b) is the modified mask decoder containing a feature fusion module with two residual blocks in parallel with attention branch. The residual block has an additional LoRA branch.

memory-centric architectural innovations: a memory encoder, memory bank, and memory attention module. Specially, the memory bank implements a FIFO queue system that stores both recently predicted frames and prompt frames, capturing spatial feature maps and object pointers to maintain temporal semantic information. To prevent cross-scene interference, we implement a prompt partitioning mechanism that selectively stores only the two most recent prompt frames alongside non-prompt predicted frames in the memory bank. The memory mechanism enables the model to effectively incorporate historical predictions and supplementary prompts into the current frame's feature processing.

Similarly, the mask decoder is modified to align with the new memory mechanism. The output of memory attention integrates high-resolution information from the hierarchical image encoder using two transposed convolution blocks as skip connections. However, the original SAM2 network directly adds high-dimensional features to embeddings, which does not effectively leverage this meta information. Prior studies [7, 23] have shown that incorporating residual blocks in the feature fusion process improves feature integration. Therefore, we introduce residual computation when fusing features at strides 4 and 8 to enhance feature refinement.

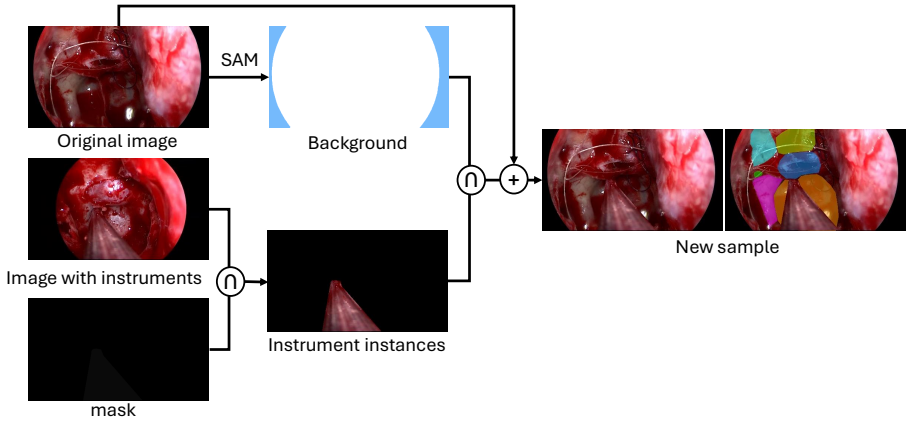

Fig. 3. Data augmentation pipeline.

Specifically, we take the high-resolution features $\mathbf{F}_{\text{high}}$ and the output from memory attention after transposed convolution $\mathbf{F}_{\text{mem}}$ as inputs to a residual block (Fig. 2(b)). The residual fusion process is formulated as:

$$\mathbf{F}_{\text{res}} = \sigma\left(\mathcal{H}(\mathbf{F}_{\text{high}}) + \mathbf{F}_{\text{mem}}\right) \tag{1}$$

where $\mathcal{H}(\cdot)$ represents a sequence of operations applied to $\mathbf{F}_{\text{high}}$, including convolution, batch normalization, and ReLU activation $\sigma(\cdot)$. Besides, we add a Low-Rank Adaptation (LoRA) branch parallel to the main convolutional path:

$$\mathbf{F}'_{\text{res}} = f\left(\mathbf{F}_{\text{res}}\right) + \alpha\mathbf{B}(\mathbf{A}\mathbf{F}_{\text{res}}) \tag{2}$$

where $\mathbf{A} \in \mathbf{R}^{r\times d}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$ are the low-rank projection matrices in the LoRA branch. r is the rank of the decomposition, typically much smaller than d to keep computations efficient. α is a scaling factor that controls the strength of the LoRA adaptation. The LoRA term $\mathbf{B}(\mathbf{A}\mathbf{F}_{\text{res}})$ provides an additional feature modulation pathway, allowing feature modulation without modifying the entire model, making it lightweight and flexible and leading to better spatial-temporal fusion.

The proposed model combines weighted focal loss, dice loss, mean-absolute-error (MAE) loss and cross-entropy (CE) loss. The over all loss function is given as follows:

$$Loss = \lambda_{focal}Loss_{focal} + \lambda_{Dice}Loss_{Dice} + \lambda_{MAE}Loss_{MAE} + \lambda_{CE}Loss_{CE} \tag{3}$$

where λ_{focal}, λ_{Dice}, λ_{MAE}, and λ_{CE} are the weights of each loss, respectively. The values are set to $20:1:1:1$, respectively, according to [16].

2.3 Data Augmentation

In our dataset, there are three larger and more distinct structures presenting in all videos: sella floor (SF), tuberculum sella (TS) and clival recess (CR). How-

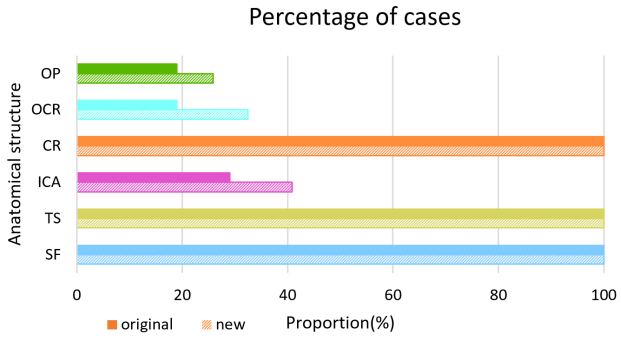

Fig. 4. Distribution of 6 anatomical structures in dataset.

ever, ICA prominence (IP), optic carotid recess (OCR), and optic prominence (OP) are significantly scarcer. Despite their low prevalence, these structures are crucial in pituitary tumor surgeries, as injuries to the internal carotid artery (ICA) and optic nerve can result in hemorrhage or vision impairment [1, 13]. As shown in Fig. 3, we implement a video reuse approach for annotating eight surgical instruments that are frequently employed during the sellar phase, as determined by expert neurosurgeons. The annotated instruments comprise: suction tube, rongeur, cutting forceps, cup forceps, bipolar electrode, freer, and scissors. We then select the cases containing ICA and OCR and superimpose surgical instruments from the additional cases into the original images in chronological order. This augmentation simulates realistic surgical scenarios, incorporating occlusions and motion artifacts caused by instruments.

3 Experiment

3.1 Dataset Description

Our dataset comprises 7,845 images extracted from 120 videos of endoscopic pituitary surgery during the sellar phase, with each frame having a resolution of either 1920×1080 or 720×576 pixels. The anatomical structures in each frame are categorized into six classes: sella floor (SF), tuberculum sella (TS), ICA prominence (IP), clival recess (CR), optic carotid recess (OCR), and optic prominence (OP). All images are annotated by specialized neurosurgeons, with a small subset labeled by researchers and later reviewed by neurosurgeons for accuracy. We first choose 100 cases and split them into 70 cases for training, 10 cases for validation, and 20 cases for testing, ensuring a balanced distribution for model evaluation. Compared to previous works [8, 19], our dataset offers more training images and higher resolution, providing a more comprehensive representation of anatomical variations. Fig. 4 illustrates the proportions of the six anatomical structures across the 100 cases in the dataset. After data augmentation, the proportions of key structures increase significantly, with ICA reaching

Table 2. Quantitative Comparison of different models

Model	mIoU	Dice						
		Mean	SF	TS	IP	CR	OCR	OP
Swin-UNet [3]	0.1872	0.2509	0.6360	0.3650	0.0114	0.4590	0.0121	0.0216
Trans-UNet [4]	0.2192	0.2847	0.7247	0.4322	0.0222	0.4806	0.0002	0.0483
DeepLabV3+ [5]	0.2085	0.2434	0.7500	0.0511	0.0002	0.5599	0.0017	0.0958
LiVOS [11]	0.4264	0.5057	0.8210	0.5851	0.2627	0.6752	0.1609	0.5293
SAM [10]	0.6090	0.7188	0.8280	0.6468	0.5988	0.7169	0.7252	0.7970
SAM-Med 2D [6]	0.6648	0.7798	0.8600	0.8403	0.6087	0.8076	0.7106	0.8514
MedSAM [12]	0.7086	0.8166	0.8641	0.8407	0.7368	0.8119	0.7792	0.8670
SAM2 [16]	0.7681	0.8397	0.9043	0.8757	0.7301	0.8667	0.8049	0.8564
Ours	0.7701	0.8559	**0.9158**	**0.8917**	0.7431	0.8826	0.8133	0.8887
Ours (+Aug)	**0.7796**	**0.8635**	**0.9158**	0.8901	**0.7821**	**0.8860**	**0.8181**	**0.8888**

40.83% and OCR 32.50%, greatly alleviating the imbalanced distribution of samples. As a result, the dataset expands to 9,331 images, with 88 cases for training and 12 for validation, and 20 cases remain unchanged for testing.

3.2 Implementation Details

We fine-tune our model based on SAM2-t pretrained weight and setting and reduce the number of prompt frames in the memory bank. The mask decoder is frozen, while all other components remain trainable. Bounding box prompts are provided for each anatomical structure every 10 frames. Input images are processed at multiple scales, with the image encoder generating 1024-resolution outputs. The implementation is based on Python 3.12.8 and PyTorch 2.5.1, running on two NVIDIA A100 Tensor Core GPUs with CUDA 12.4. Training is conducted for 40 epochs, and the best model is obtained using the AdamW optimizer($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a base learning rate of 5.0×10^{-6}.

3.3 Results

To evaluate the effectiveness of our model, we train unprompted image models Swin-UNet [3], Trans-UNet [4], DeepLabV3+ [5], video segmentation model LiVOS [11], fully-prompted model SAM-Med 2D [6], MedSAM [12] and the original SAM [10] and SAM2 [16] and compare their performance. As shown in Table 2, our F2PASeg achieves a mean Dice score of 85.59%, which is 1.62% higher than the original SAM2 and 4.69% higher than the fully-prompted MedSAM. Due to the complexity of the scene, the unprompted models perform poorly and can only detect three more obvious structures. Fig. 5 provides a visual comparison of segmentation results. F2PASeg demonstrates superior segmentation in dynamically changing regions, particularly those affected by bleeding or instrument occlusion during surgery. In particular, in the second and fourth columns of Table 2, TS (top of frame) and CR (bottom of frame) are often affected by camera pans. The feature fusion module improves segmentation continuity, ensuring that predictions remain more stable and aligned with the prompt frames. Moreover, F2PASeg achieves 28.57 FPS, which is 2.3 times higher than that

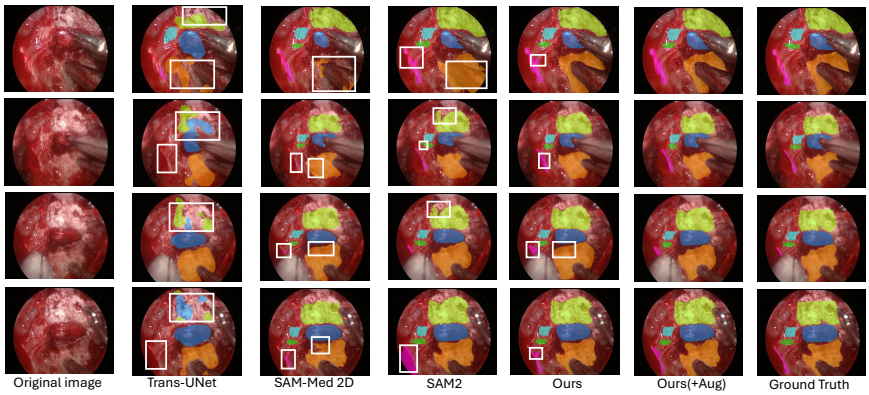

Fig. 5. visualization result

Table 3. Ablation Studies Results

Feature Fusion Module	Data Augmentation	mDice	mIoU
-	-	0.8397	0.7681
-	✓	0.8531	0.7697
✓	-	0.8559	0.7701
✓	✓	**0.8635**	**0.7796**

of SAM-Med 2D, and basically meets the requirement of intraoperative real-time segmentation. With the LoRA module, the number of training parameters decreased from 39.0M to 34.8M.

Ablation Studies To verify the validity of our designed model and data augmentation strategy, we conduct the ablabtion studies. The detailed results with different configurations are listed in Table 3. These results indicate that our model enhances segmentation performance by effectively modeling the relationships between anatomical structures within the feature fusion module. Additionally, the third row of Table 2 presents detailed results, revealing segmentation accuracy improvements of 3.90% for ICA and 0.48% for OCR compared to the previous model. Notably, F2PASeg reduces incorrect segmentation of surgical instruments relative to the original model. Additionally, our data-augmented model further enhances the segmentation accuracy of ICA, reinforcing the efficiency of our spatial feature extraction method.

4 Conclusion

In this paper, we addressed pituitary anatomy segmentation during the sellar phase of pituitary surgeries. We introduced a large-scale dataset, Pituitary Anatomy Segmentation (PAS), comprising 7,845 high-resolution, temporally coherent images from 120 surgeries. Each image has been meticulously annotated

by expert neurosurgeons. We proposed F2PASeg, an efficient architecture designed to explicitly model relationships between anatomical structures in endoscopic surgery. Our method achieved a mean Dice score of 86.35%. The segmentation accuracy for carotid arteries, a critical structure for surgical safety, was notably enhanced. This improvement provides early warnings of high-risk regions, assisting surgeons in their procedures. In addition, F2PASeg meets the real-time processing requirements for video-based intraoperative applications. This ensures seamless integration into surgical workflows, providing live, high-accuracy anatomical segmentation during endoscopic procedures. As future work, we plan to explore adaptive temporal modeling to further enhance segmentation robustness in long video sequences.

Acknowledgments. The research in this paper was funded by Inno HK program.

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Arnold, M.A., Barbero, J.M.R., Pradilla, G., Wise, S.K.: Pituitary gland surgical emergencies: the role of endoscopic intervention. Otolaryngologic Clinics of North America **55**(2), 397–410 (2022)
2. Bonneville, J.F.: Magnetic resonance imaging of pituitary tumors. Imaging in Endocrine Disorders **45**, 97–120 (2016)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Proceedings of European Conference on Computer Vision. pp. 205–218. Springer (2022)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
6. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
7. Chobola, T., Müller, G., Dausmann, V., Theileis, A., Taucher, J., Huisken, J., Peng, T.: Lucyd: A feature-driven richardson-lucy deconvolution network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 656–665. Springer (2023)
8. Das, A., Khan, D.Z., Williams, S.C., Hanrahan, J.G., Borg, A., Dorward, N.L., Bano, S., Marcus, H.J., Stoyanov, D.: A multi-task network for anatomy identification in endoscopic pituitary surgery. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 472–482. Springer (2023)
9. Hao, S., Zhou, Y., Guo, Y.: A brief survey on semantic segmentation with deep learning. Neurocomputing **406**, 302–321 (2020)

10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
11. Liu, Q., Wang, J., Yang, Z., Li, L., Lin, K., Niethammer, M., Wang, L.: Livos: Light video object segmentation with gated linear matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8668–8678 (2025)
12. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
13. Madani, A., Namazi, B., Altieri, M.S., Hashimoto, D.A., Rivera, A.M., Pucher, P.H., Navarrete-Welton, A., Sankaranarayanan, G., Brunt, L.M., Okrainec, A., et al.: Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Annals of surgery **276**(2), 363–369 (2022)
14. Marcus, H.J., Khan, D.Z., Borg, A., Buchfelder, M., Cetas, J.S., Collins, J.W., Dorward, N.L., Fleseriu, M., Gurnell, M., Javadpour, M., et al.: Pituitary society expert delphi consensus: operative workflow in endoscopic transsphenoidal pituitary adenoma resection. Pituitary **24**(6), 839–853 (2021)
15. Patel, C.R., Fernandez-Miranda, J.C., Wang, W.H., Wang, E.W.: Skull base anatomy. Otolaryngologic Clinics of North America **49**(1), 9–20 (2016)
16. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
18. Sarwin, G., Carretta, A., Staartjes, V., Zoli, M., Mazzatenta, D., Regli, L., Serra, C., Konukoglu, E.: Live image-based neurosurgical guidance and roadmap generation using unsupervised embedding. In: International Conference on Information Processing in Medical Imaging. pp. 107–118. Springer (2023)
19. Staartjes, V.E., Volokitin, A., Regli, L., Konukoglu, E., Serra, C.: Machine vision for real-time intraoperative anatomic guidance: a proof-of-concept study in endoscopic pituitary surgery. Operative Neurosurgery **21**(4), 242–247 (2021)
20. Van Furth, W.R., De Vries, F., Lobatto, D.J., Kleijwegt, M.C., Schutte, P.J., Pereira, A.M., Biermasz, N.R., Verstegen, M.J.: Endoscopic surgery for pituitary tumors. Endocrinology and Metabolism Clinics **49**(3), 487–503 (2020)
21. Wang, F., Zhou, T., Wei, S., Meng, X., Zhang, J., Hou, Y., Sun, G.: Endoscopic endonasal transsphenoidal surgery of 1,166 pituitary adenomas. Surgical endoscopy **29**, 1270–1280 (2015)
22. Wu, Z., Lau, C.Y., Zhou, Q., Wu, J., Wang, Y., Liu, Q., Lei, Z., Liu, H.: Surgivisor: Transformer-based semi-supervised instrument segmentation for endoscopic surgery. Biomedical Signal Processing and Control **87**, 105434 (2024)
23. Zhao, M., Kang, M., Tang, B., Pecht, M.: Deep residual networks with dynamically weighted wavelet coefficients for fault diagnosis of planetary gearboxes. IEEE Transactions on Industrial Electronics **65**(5), 4290–4300 (2017)