This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

See in Depth: Training-Free Surgical Scene Segmentation with Monocular Depth Priors

Kunyi Yang[1], Qingyu Wang[1,2], Cheng Yuan[1][0000−0001−8083−3026], Alaa Maalouf[3], and Yutong Ban[1⋆]

[1] Shanghai Jiao Tong University, Shanghai, China
yban@sjtu.edu.cn
[2] University of Michigan, Ann Arbor, USA
[3] University of Haifa, Haifa, Israel

Abstract. Dense manual annotations remain a bottleneck for scaling surgical scene understanding to large video collections. Given annotated reference frames, we introduce Depth-guided Surgical Scene Segmentation (DepSeg), a training-free framework that propagates their semantic information to unlabeled videos without task-specific model training. DepSeg uses monocular depth as a geometric prior to generate point prompts for SAM2 and classifies the resulting masks through feature-level template matching. Evaluated on CholecSeg8k (72.11% mIoU) and EndoVis 2018 (37.58% mIoU), DepSeg outperforms standard automatic baselines and exceeds a fully supervised model in the one-video setting. Although a gap remains relative to fully supervised methods, this depth-aware pipeline provides masks sufficient to support various downstream tasks.

Keywords: Surgical Scene Segmentation · Laparoscopic Surgery · Surgical Scene Understanding

1 Introduction

Accurate pixel-wise segmentation of surgical scenes is important for surgical video analysis and dataset curation. However, hospitals may have a small set of annotated reference frames alongside a much larger collection of unlabeled archived videos. We consider propagating the semantic information in these references to the unlabeled collection without task-specific model training. DepSeg uses annotated references to construct its non-parametric feature bank. Then, it propagates the limited reference annotations to those unlabeled videos at scale with pre-trained vision foundation models. Moreover, SurgWorld[7] shows that even imperfect segmentation masks, when generated at scale, provide valuable utility for various downstream surgical applications.

Recently, vision foundation models have developed rapidly to offer powerful tools. SAM [9] and SAM2 [18] provide strong, class-agnostic masks and have

⋆ Corresponding author

been adapted for medical data via Surgical SAM2 [12] and MedSAM2 [13]. Benchmark [25] shows that SAM2 is capable of generating high-quality masks in surgical scenarios when provided with proper prompts. Concurrently, DINO self-supervised encoders [4, 16, 20] produce rich features that transfer well across domains. Recent training-free methods combine SAM masks with DINO descriptors [15] or use feature memory banks [5] to automatically generate instance masks without fine-tuning. While fully supervised networks have developed specialized spatial and temporal architectures to handle complex surgical textures [23, 24], unlocking the full potential of frozen foundation models requires a reliable mechanism to automatically generate these proper spatial prompts.

In laparoscopic scenes, instruments and anatomical layers naturally appear at distinct depth levels, forming a structured spatial layout. Pre-trained monocular depth models like DepthAnythingv2 [22] and DepthAnything3 [11] offer a direct way to reliably capture these relative spatial relationships. These geometric cues provide the ideal foundation for automatically generating spatial prompts to guide the segmentation process. By extracting this depth information as a lightweight geometric prior, we enable segmentation models to use depth differences to identify boundaries, effectively dividing the surgical area into distinct objects even when the surfaces look the same.

We propose DepSeg (Depth-guided Surgical Scene Segmentation), a training-free framework for automated surgical scene parsing. DepSeg combines depth cues with a frozen SAM2 model and pre-trained feature extractor (DINOv3). We extract 3D physical cues from monocular depth estimates to generate depth-guided point prompts, which SAM2 converts into class-agnostic masks. Each mask is then described by a feature vector and classified via feature-level template matching. Our main contributions are as follows:

1. We propose DepSeg, a training-free framework that uses frozen foundation models for surgical scene segmentation. Given annotated reference frames, it propagates semantic masks without task-specific model training or retraining when new references are added.
2. We introduce a depth-guided prompting strategy for laparoscopic surgical scenes. By using monocular depth as a geometric prior that automatically generates spatial prompts, our method enables foundation models to separate distinct objects despite uniform surfaces.
3. Experiments on the CholecSeg8k and EndoVis 2018 datasets show that our method achieves robust segmentation of both instruments and tissues. While a gap remains relative to fully-supervised methods, the segmentation quality is already sufficient to support different downstream tasks.

2 Method

DepSeg consists of two phases: Surgical Feature Bank Construction and Online Inference (Fig. 1). The construction phase builds a non-parametric feature bank from annotated reference frames using a frozen pre-trained feature extractor.

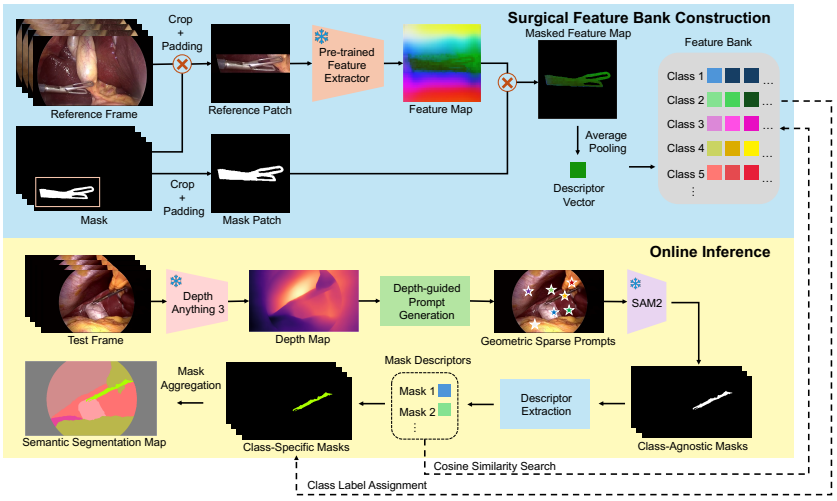


Fig. 1. Overview of the training-free DepSeg framework. (Top) Feature Bank Construction: a pre-trained feature extractor extracts descriptor vectors from masked reference patches to build a surgical feature bank. (Bottom) Online Inference: Depth Anything 3 generates depth maps, which are then used to extract geometric sparse prompts. These guide SAM2 to produce class-agnostic masks. Extracted mask descriptors are classified via cosine similarity against the feature bank and aggregated into the final semantic map. (Snowflake icons denote frozen models).

During inference, this bank assigns semantic labels to masks from new videos. Depth-guided prompts from each test frame direct SAM2 to generate class-agnostic masks, which are classified using cosine similarity against the feature bank and aggregated into the final pixel-wise segmentation map.

2.1 Problem formulation

Let $\Omega = \{1, \ldots, H\} \times \{1, \ldots, W\}$ denote the image domain and $x \in \mathbb{R}^{H \times W \times 3}$ a single RGB laparoscopic frame. We consider a finite set of foreground classes $\mathcal{C}$ that includes instrument and tissue categories, and define the pixelwise label space $\mathcal{Y} = \{0\} \cup \mathcal{C}$, where 0 denotes background. The objective is to predict a label for every pixel:

$$F : \mathbb{R}^{H \times W \times 3} \to \mathcal{Y}^{H \times W}, \tag{1}$$

where $y = F(x)$ is the final segmentation map.

2.2 Surgical Feature Bank Construction

This offline stage aims to generate a feature bank of class-representative descriptors. The process uses a labeled training set, denoted as $\{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$, where N is the total number of training samples, $x^{(n)}$ is the n-th reference frame, and

$y^{(n)}$ is its corresponding pixel-wise ground truth label map. For each semantic class c from the set of all possible labels $\mathcal{Y}$, a binary mask $M_{n,c}$ is created from the ground truth map $y^{(n)}$. This mask identifies all pixels belonging to class c:

$$M_{n,c}(p) = \mathbf{1}[y^{(n)}(p) = c], \quad p \in \Omega, \tag{2}$$

where p denotes a pixel coordinate within the image spatial domain Ω, $\mathbf{1}[\cdot]$ is the indicator function, which is 1 if the condition inside is true and 0 otherwise. To preserve the geometric proportions of objects, a bounding box is extracted around the mask $M_{n,c}$, expanded by 15%, padded with black color to form a square, and resized to 224×224. This letterboxed image crop is then processed by a frozen pre-trained feature extractor (DINOv3) to produce a token grid $T \in \mathbb{R}^{H' \times W' \times D}$. The corresponding letterboxed mask is similarly downsampled to the token grid's resolution ($H' \times W'$) to create a new mask $M'_{n,c}$. A mask-aware pooled descriptor $h_{n,c}$ is then computed by averaging the token vectors over the spatial locations (u, v) covered by the downsampled mask:

$$h_{n,c} = \frac{1}{Z_{n,c}} \sum_{u,v} M'_{n,c}(u, v)\, T(u, v, :) \in \mathbb{R}^D, \quad Z_{n,c} = \sum_{u,v} M'_{n,c}(u, v) \tag{3}$$

where $T(u, v, :)$ refers to the D-dimensional feature vector at grid location (u, v), and the normalization factor $Z_{n,c}$ is the area of the downsampled mask. The resulting descriptor $h_{n,c}$ is then L2-normalized. Finally, all descriptors for a given class c are collected from the entire training set to form a feature bank, $\mathcal{T}_c = \{t_{c,\ell}\}_{\ell=1}^{L_c}$. Here, each $t_{c,\ell}$ is a normalized descriptor vector. This bank is saved and loaded during the inference phase.

2.3 Online Inference

The online inference phase processes each test frame sequentially to produce the final semantic segmentation map. This phase consists of three main steps.

Depth-guided Prompt Generation This stage takes a test frame $x \in \mathbb{R}^{H \times W \times 3}$ to generate sparse, high-quality point prompts. First, a pretrained monocular depth estimator, DepthAnything3 [11], produces a dense depth field $d \in \mathbb{R}^{H \times W}$. From this, a relative depth map d_{rel} is calculated by subtracting the median depth value of the image. Using d_{rel} and x, we apply a multi-heuristic proposal operator that aggregates parallel results from K-means clustering [14], Canny edge detection [3], Otsu thresholding [17], and a watershed transform [21] to generate initial region proposals $\mathcal{R} = \{R_m\}_{m=1}^{M}$, where $R_m \subseteq \Omega$. These proposals are regularized using morphological operators to yield cleaned regions $\tilde{R}_m$. For each cleaned region $\tilde{R}_m$, a distance transform is computed:

$$D_m(p) = \text{dist}(p, \partial\tilde{R}_m), \qquad p \in \Omega, \tag{4}$$

where $\partial\tilde{R}_m$ is the region boundary. A set of three geometric prompts per region, S_m, is obtained by detecting local maxima of the distance map $D_m(p)$, subject to constraints on minimum distance and border exclusion.

SAM-based Mask Generation This stage uses the point prompts $\{S_m\}$ to produce a set of class-agnostic object masks. A frozen SAM2[18] image predictor takes the image x and the prompts S_m for each proposed region to generate initial binary masks, which are then cleaned with a morphological opening operator. Masks with an area smaller than a minimum threshold are discarded.

To resolve overlaps, masks are merged into a priority label map L by sorting them by ascending area. Each pixel is assigned the index of the smallest mask covering it, ensuring the preservation of fine details. The map is then decomposed into spatially connected components for each unique index. Each component is treated as an individual binary mask, forming the final refined set: $\widehat{\mathcal{M}} = \{M_j \mid j = 1, \ldots, J\}$, where J is the number of final masks.

Descriptor Extraction and Class Label Assignment In this final stage, each refined mask $M_j \in \widehat{\mathcal{M}}$ undergoes descriptor extraction, class assignment, and aggregation to produce the semantic segmentation map. First, a descriptor h_j is computed using the mask-aware pooling process from the construction stage. The mask M_j is downsampled to the token grid resolution to create M'_j, and the descriptor is L2-normalized:

$$h_j = \text{norm}\left(\frac{1}{Z_j}\sum_{u,v} M'_j(u,v)\, T(u,v,:)\right) \in \mathbb{R}^D, \quad Z_j = \sum_{u,v} M'_j(u,v) \tag{5}$$

where Z_j represents the area of the downsampled mask M'_j.

Next, each h_j is compared against pre-computed feature banks $\{\mathcal{T}_c\}$. The score $S_{j,c}$ is calculated by summing the top-k cosine similarities between h_j and bank descriptors $t_{c,\ell}$, with $k = 10$:

$$s_{j,c,\ell} = h_j^\top t_{c,\ell}, \quad S_{j,c} = \sum_{\ell \in \text{TopK}_\ell(s_{j,c,\ell})} s_{j,c,\ell}. \tag{6}$$

To ensure temporal stability and mitigate flickering, an Exponential Moving Average (EMA) is applied. If a mask M_j at frame t shares an Intersection over Union (IoU) > 0.5 with a mask from frame $t-1$, the score is smoothed using weight $\alpha = 0.7$:

$$S_{j,c}^{(t)} = \alpha S_{j,c} + (1-\alpha) S_{j,c}^{(t-1)}. \tag{7}$$

The mask M_j is then assigned the class $\hat{c}_j$ with the highest smoothed score:

$$\hat{c}_j = \arg\max_{c \in \mathcal{Y}} S_{j,c}^{(t)}. \tag{8}$$

Finally, the labeled masks $\{(M_j, \hat{c}_j)\}$ are sorted by ascending area to generate the final semantic map y. In this small-area-first merge strategy, each mask assigns its label only to previously unlabeled pixels, preserving the details of small objects in the final prediction $y = F(x)$.

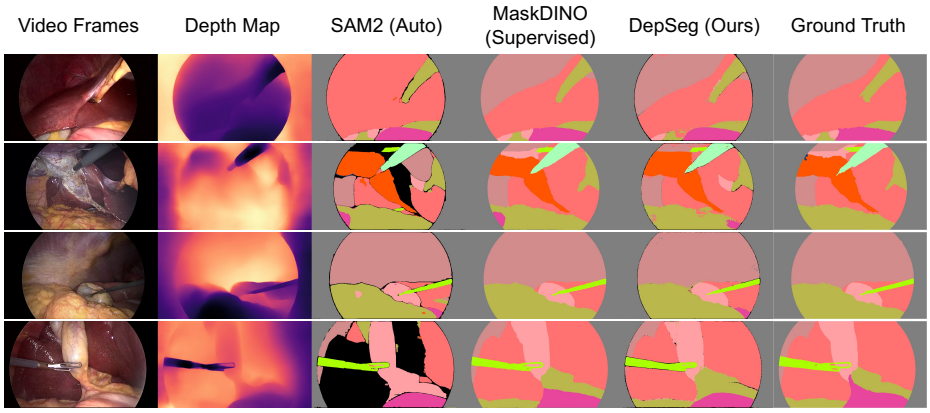


Fig. 2. Qualitative semantic segmentation results. From left to right: original RGB images, estimated depth maps, segmentation results from SAM2 (Auto), MaskDINO, our proposed DepSeg, and the Groundtruth.

3 Experiments and Results

3.1 Dataset

We evaluate our approach on two datasets to demonstrate generalizability across different surgical domains. The CholecSeg8k dataset [8] provides 8,080 laparoscopic cholecystectomy frames from 17 video clips, which we split into 6,400 for training and 1,680 for validation. Following [19], we adopt an 8-class categorization for the anatomical structures and instruments. The EndoVis 2018 dataset [1] annotated by [6] contains 2,235 robotic-surgery frames from 15 sequences. It features semantic masks for the background and seven instrument categories. We follow its official split, using 1,639 frames (11 sequences) for training and 596 frames (4 sequences) for validation.

All models are used as frozen backbones. Since DepthAnything3, SAM2, and DINOv3 are applied sequentially, DepSeg incurs non-trivial overhead and is intended for offline processing of archived surgical videos rather than real-time intraoperative deployment.

3.2 Results and Discussion

Results on CholecSeg8k Table 1 reports the per-class Intersection over Union and the mean metric, mIoU, on the CholecSeg8k dataset. Fully supervised models, such as EndoViT and MaskDINO, establish the performance upper bound with mIoU scores of 86.86% and 87.52%. To contextualize our training-free framework, we compare DepSeg against two foundational baselines: SAM2 (Auto), which uses the default automatic mask generator, and SAM2 (GT). To construct SAM2 (GT), we derive point prompts directly from the ground truth

Table 1. Quantitative results on CholecSeg8k. Connective Tissue, Blood, Cystic Duct, Hepatic Vein, and Liver Ligament are merged into Misc., while Grasping Forceps and L-hook Electrocautery are merged into Inst.. All metrics are reported as percentages (%). (mIoU: mean IoU, Bkg.: Background, AW: Abdominal Wall, Liv.: Liver, GI: Gastrointestinal Tract, Fat: Fat, GB: Gallbladder, Misc.: Miscellaneous, Inst.: Instruments)

Method	Superv.	mIoU	Bkg.	AW	Liv.	GI	Fat	GB	Misc.	Inst.
Fully Supervised Methods										
EndoViT [2]	Full	86.86	99.24	93.43	93.51	82.58	91.58	87.10	59.95	87.47
MaskDINO [10]	Full	87.52	99.01	93.88	93.80	84.06	91.78	89.94	60.21	87.46
Ground Truth-Based Methods										
SAM2 (GT)	None	80.57	97.68	87.37	86.30	71.22	84.90	82.61	54.75	79.70
Training-Free Methods										
SAM2 (Auto)	None	70.37	96.81	79.73	69.13	47.97	76.40	73.24	39.12	80.55
DepSeg (Ours)	None	72.11	97.18	80.28	81.00	46.50	77.59	77.59	43.14	73.58

annotations. Specifically, we first split each ground truth mask into its distinct connected components. For every component, we compute the distance to the boundaries and extract three points corresponding to the local maximal distance (i.e., the points deepest within the region). These three points are then used as positive point prompts for SAM2. All subsequent steps of SAM2 (Auto) and SAM2 (GT), including feature extraction, and template matching, remain identical to our DepSeg pipeline. Our proposed DepSeg achieves an mIoU of 72.11%, consistently outperforming the SAM2 (Auto) baseline at 70.37%. This 1.74% improvement demonstrates that our depth-guided geometric prompts are more effective than the default grid-based automatic generator. The results are visually presented in Fig. 2, which provides a comparison of the segmentation outputs. While SAM2 (Auto) frequently under-segments or misses objects due to a lack of geometric guidance, DepSeg effectively uses the depth prior to ensure SAM2 focuses on the target structures. Although DepSeg does not match the peak performance of fully supervised methods, it demonstrates that a simple, training-free framework provides sufficiently accurate segmentation that supports different downstream tasks.

Results on EndoVis 2018 To validate the robustness of our approach, we report performance on the EndoVis 2018 dataset in Table 2. DepSeg achieves an mIoU of 37.58%, surpassing the SAM2 (Auto) baseline at 34.31%. DepSeg also outperforms the overall mIoU of the SAM2 (GT) baseline, achieving 37.58% compared to 34.93%, which is driven by a 29.51% improvement in the background class (92.21% versus 62.70%). In the EndoVis 2018 dataset, the background represents the entire surgical scene, including all tissues, organs and anatomical structures not classified as instruments. This shows that prompts derived from depth are more effective at capturing the surgical scene than simple spatial points taken from the ground truth.

Table 2. Quantitative evaluation on the EndoVis 2018 dataset. All metrics are reported as percentages (%). (mIoU: mean IoU, BG: Background, BF: Bipolar Forceps, PF: Prograsp Forceps, LND: Large Needle Driver, MCS: Monopolar Curved Scissors, UP: Ultrasound Probe, SI: Suction Instrument, CA: Clip Applier)

Method	Superv.	mIoU	BG	BF	PF	LND	MCS	UP	SI	CA
Fully Supervised Methods										
EndoViT [2]	Full	43.76	97.07	86.44	19.92	0.66	93.40	12.66	39.91	0.00
MaskDINO [10]	Full	42.52	98.26	78.62	39.42	11.02	93.38	0.01	19.43	0.00
Ground Truth-Based Methods										
SAM2 (GT)	None	34.93	62.70	53.36	12.20	13.47	89.45	3.49	44.76	0.00
Training-Free Methods										
SAM2 (Auto)	None	34.31	87.13	53.81	8.39	9.66	83.47	2.56	29.45	0.00
DepSeg (Ours)	None	37.58	92.21	59.11	9.85	11.72	86.88	1.26	39.60	0.00

Table 3. Quantitative evaluation across annotated-video counts on CholecSeg8k.

Method	Count	mIoU	Bkg.	AW	Liv.	GI	Fat	GB	Misc.	Inst.
Fully Supervised Methods										
MaskDINO [10]	1	46.43	97.10	67.01	71.19	2.97	56.80	51.70	0.00	24.66
	2	53.68	98.76	79.36	79.42	20.73	69.80	52.93	0.00	28.43
	4	62.35	98.97	84.13	85.40	54.64	79.50	66.75	0.00	29.43
	13	80.32	99.03	93.00	91.82	75.06	86.44	82.68	30.49	84.02
	Full	87.52	99.01	93.88	93.80	84.06	91.78	89.94	60.21	87.46
Training-Free Methods										
DepSeg(Ours)	1	48.78	97.46	65.12	69.75	26.96	54.82	45.80	0.00	30.35
	2	49.75	94.41	68.14	70.76	24.78	61.99	47.78	0.00	30.15
	4	55.37	94.65	73.12	75.27	26.33	65.83	56.92	0.00	50.83
	13	65.87	96.55	78.27	78.70	41.88	72.29	72.81	21.60	64.86
	Full	72.11	97.18	80.28	81.00	46.50	77.59	77.59	43.14	73.58

Annotated-Video Count Ablation For DepSeg, annotated reference videos populate the non-parametric feature bank. Table 3 compares DepSeg with fully supervised MaskDINO as the number of annotated videos varies. With one video, DepSeg achieves 48.78% mIoU versus 46.43% for MaskDINO. This indicates that DepSeg is competitive in the one-video setting.

4 Conclusion

We presented DepSeg, a training-free framework for surgical segmentation that combines depth-guided prompts for SAM2 with feature-level template matching. Given annotated reference frames, DepSeg propagates semantic labels to surgical videos without task-specific model training. DepSeg improves over direct

SAM2 baselines on both the CholecSeg8k and EndoVis 2018 datasets. Requiring only forward passes on frozen foundation models, DepSeg provides a robust and extendable solution for surgical scene understanding, and is intended for offline annotation propagation. Future work will explore multi-view geometry and generated masks for downstream surgical tasks.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Batić, D., Holm, F., Özsoy, E., Czempiel, T., Navab, N.: Endovit: pretraining vision transformers on a large collection of endoscopic images. International Journal of Computer Assisted Radiology and Surgery **19**(6), 1085–1091 (2024)
3. Canny, J.: A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence (6), 679–698 (2009)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
5. Espinosa, M., Yang, C., Ericsson, L., McDonagh, S., Crowley, E.J.: No time to train! training-free reference-based instance segmentation. arXiv preprint arXiv:2507.02798 (2025)
6. González, C., Bravo-Sánchez, L., Arbelaez, P.: Isinet: an instance-based approach for surgical instrument segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 595–605. Springer (2020)
7. He, Y., Guo, P., Xu, M., Li, Z., Myronenko, A., Imans, D., Liu, B., Yang, D., Gu, M., Ji, Y., Jin, Y., Zhao, R., Shen, B., Xu, D.: Surgworld: Learning surgical robot policies from videos via world modeling. arXiv preprint arXiv:2512.23162 (Dec 2025). https://doi.org/10.48550/arXiv.2512.23162, `https://arxiv.org/abs/2512.23162`
8. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. CoRR **abs/2012.12453** (2020), `http://dblp.uni-trier.de/db/journals/corr/corr2012.html#abs-2012-12453`
9. 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)
10. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3041–3050 (2023)
11. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)

12. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. arXiv preprint arXiv:2408.07931 (2024)
13. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025)
14. MacQueen, J.: Multivariate observations. In: Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability. vol. 1, pp. 281–297 (1967)
15. Nguyen, V.N., Groueix, T., Ponimatkin, G., Lepetit, V., Hodan, T.: Cnos: A strong baseline for cad-based novel object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2134–2140 (2023)
16. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
17. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica **11**(285-296), 23–27 (1975)
18. 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)
19. Silva, B., Oliveira, B., Morais, P., Buschle, L., Correia-Pinto, J., Lima, E., Vilaça, J.L.: Analysis of current deep learning networks for semantic segmentation of anatomical structures in laparoscopic surgery. In: 2022 44th annual international conference of the IEEE engineering in medicine & biology society (EMBC). pp. 3502–3505. IEEE (2022)
20. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
21. Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis & Machine Intelligence **13**(06), 583–598 (1991)
22. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
23. Yuan, C., Ban, Y.: Surgical scene segmentation by transformer with asymmetric feature enhancement. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
24. Yuan, C., Ban, Y.: Temporal propagation of asymmetric feature pyramid for surgical scene segmentation. In: International Workshop on Collaborative Intelligence and Autonomy in Image-Guided Surgery. pp. 187–196. Springer (2025)
25. Yuan, C., Jiang, J., Yang, K., Wu, L., Wang, R., Meng, Z., Ping, H., Xu, Z., Zhou, Y., Song, W., et al.: Is segment anything model 2 all you need for surgery video segmentation? a systematic evaluation. arXiv preprint arXiv:2501.00525 (2024)