Evaluating the Prompt-to-Automation Gap: A Comprehensive Benchmark for Laparoscopic Endometriosis Segmentation

Jasmin Arjomandi[1[0009−0008−6801−8676]], Christian Kunz[1[0000−0002−0571−901X]], Xingjian Kang[2[0009−0005−0590−1734]], Katharina Breininger[1,2[0000−0001−7600−5869]], and Franziska Mathis-Ullrich[1[0000−0001−5239−5305]]

[1] Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany
[2] Center for AI and Data Science (CAIDAS), Julius-Maximilians Universität Würzburg, Würzburg, Germany
jasmin.arjomandi@fau.de

Abstract. Accurate segmentation of endometriosis lesions in laparoscopic images remains challenging due to subtle boundaries, heterogeneous appearances and locations, and limited annotated data. While SAM-based foundation models enable flexible promptable segmentation, their value for fully automatic AI-assisted endometriosis surgery remains unclear. We present a cross-complementary dataset benchmark of supervised, SAM-based promptable, and hybrid automatic approaches for laparoscopic endometriosis lesion segmentation. Experiments were conducted on ENID, dominated by implant-like lesions, and GLENDA-clean, containing more heterogeneous lesion appearances and anatomical locations. We evaluated SAM2, MedSAM, SAM-Med2D, and SurgiSAM2 without fine-tuning under four different controlled prompt settings. Box prompts were most reliable, increasing mean Dice over point prompts by 0.179 on ENID and 0.410 on GLENDA-clean, while additional positive or negative points did not significantly improve performance. SAM2 achieved the highest Dice on ENID, whereas SurgiSAM2 performed best on GLENDA-clean. We then trained DeepLabV3+, SegFormer, YOLO11s-seg, U-Net++ and nnU-Net v2, as fully automatic baselines, with SegFormer showing the strongest overall cross-dataset performance. Finally, as a hybrid experiment, we evaluated SegFormer-to-SurgiSAM2 automatic refinement using predicted lesion candidates as prompts. Fallback acceptance reduced degradation compared with direct refinement, but no hybrid variant outperformed SegFormer alone.
Our code and benchmark is available at: https://github.com/JArjomandi/Auto_Endometriosis_Segmentation

Keywords: Endometriosis lesion segmentation · Laparoscopic surgery · Promptable foundation models · Hybrid automatic segmentation

1 Introduction and Related Works

Endometriosis is a common, painful, and heterogeneous gynecological disease in women of reproductive age, in which endometrial-like lesions occur outside the uterus across diverse pelvic and abdominal locations, with laparoscopy remaining central to diagnosis and surgical treatment planning [12,13]. Reliable visual interpretation is challenging, since lesions vary in size, color, morphology, location, and contrast against surrounding tissue, and may be obscured by blood, smoke, specular reflections, deformation, or surgical manipulation [13,1]. Automated lesion segmentation could support AI-assisted laparoscopic guidance through frame-wise localization, extent estimation, documentation, tracking, and decision support, but requires dense masks rather than only image-level labels or bounding boxes [1,19,2]. However, dense endometriosis annotation is clinically demanding, time-consuming, and expert-dependent, especially for small, subtle, multifocal, or ambiguous lesions, contributing to the scarcity of high-quality gynecologic laparoscopic datasets [12,18].

Compared with instrument, anatomy, and surgical scene segmentation, gynecologic laparoscopic pathology segmentation remains underexplored, and endometriosis -specific segmentation frameworks have rarely been evaluated systematically across datasets [11,9,13,1]. Existing endometriosis AI studies have mainly addressed detection, localization, or clinical recognition, using Faster R-CNN, Mask R-CNN, YOLO-style detectors, or proof-of-concept clinical recognition systems, but do not establish robust pixel-wise segmentation across modern supervised, promptable, and hybrid paradigms under heterogeneous laparoscopic lesion appearance and dataset shift [13,1,19]. Endometriosis segmentation has also been explored in ultrasound and MRI, but these modalities differ substantially from 2D laparoscopic surface lesions in appearance, annotation, and deployment requirements [20,5].

Task-specific models such as DeepLabV3+, U-Net++, nnU-Net, and SegFormer provide strong biomedical segmentation baselines, but their performance on small, irregular, and visually ambiguous laparoscopic endometriosis lesions remains insufficiently characterized [3,26,7,24,11]. SAM-based models offer an alternative promptable paradigm, with SAM/SAM2 providing general segmentation and MedSAM, SAM-Med2D, and SurgiSAM2 targeting medical or surgical domains [10,21,16,4,9]. Yet recent evaluations show that SAM performance is highly prompt-dependent, with box prompts often outperforming point prompts and imperfect automatic prompts limiting deployment [17,15,25]. Hybrid automatic prompting and SAM refinement have been explored for surgical instruments, ultrasound, skin lesions, and polyps, but their value for subtle laparoscopic endometriosis lesions remains unclear [22,15,6,14,25].

In this work, we evaluate the transition from promptable to fully automatic endometriosis lesion segmentation in laparoscopic images. We benchmark supervised, SAM-based promptable, and hybrid automatic approaches across ENID and GLENDA datasets [12,13], evaluate controlled prompt settings, and as an ablation experiment, test SegFormer-to-SurgiSAM2 refinement with fallback acceptance. To our knowledge, this is the first cross-dataset benchmark jointly

evaluating supervised, promptable, and hybrid automatic segmentation for laparoscopic endometriosis lesions, explicitly quantifying the gap between interactive prompt-based masks and deployable automatic segmentation, for future AI-assisted endometriosis surgery guidance.

2 Methods and Experiments

2.1 Datasets

We evaluated all models on two complementary laparoscopic endometriosis datasets. ENID contains 160 frames from 108 gynecologic laparoscopy cases with 358 region-based annotations of dark endometrial implants [13], derived from GLENDA, a broader laparoscopy endometriosis dataset extracted from over 400 surgical videos, with version 1.5 providing 373 annotated frames, 628 annotations, and 13,438 unannotated non-pathology frames across peritoneal, ovarian, uterine, deep infiltrating, and no-pathology categories [12]. Since GLENDA includes both hand-drawn masks and box-like region annotations, we created GLENDA-clean by excluding bounding-box masks and annotation artifacts (328 frames). Primary results are reported on GLENDA-clean, with original GLENDA retained only as an annotation-quality sensitivity analysis.

All data were standardized as images with corresponding converted binary masks. Case-aware training, validation, and test splits avoided leakage between patient cases or video-derived image groups. Final split sizes were 96/32/32 for ENID and 210/52/ 64 for GLENDA-clean. Validation sets were used for model selection, threshold calibration, and early stopping; held-out test sets were used for final comparison.

2.2 Models

We evaluated two model groups: SAM-based promptable foundation models, fully automatic supervised models, as well as a hybrid for automatic refinement. SAM2 was included as a general promptable image/video segmentation foundation model [21]; MedSAM and SAM-Med2D as medical-domain SAM variants [16,4]; and SurgiSAM2 as a surgical-domain SAM2-based model [9]. These models were evaluated without task-specific fine-tuning to assess out-of-the-box promptable segmentation performance.

The supervised benchmark included representative task-specific segmentation families. DeepLabV3+ was used as a CNN encoder–decoder with multi-scale context aggregation [3]; SegFormer as a lightweight Transformer-based semantic segmentation model [24]; YOLO11-seg as an efficient YOLO-based detection/segmentation-style model [23,8]; also U-Net++ as a nested U-Net with dense skip connections [26]; and nnU-Net v2 as a self-configuring biomedical segmentation baseline [7]. These models were selected to cover common CNN, Transformer, self-configuring, and real-time segmentation paradigms while avoiding unnecessarily heavy architectures, relevant for future laparoscopic video analysis.

2.3 Implementation

We designed the study as a three-stage benchmark to evaluate the transition from promptable to fully automatic endometriosis lesion segmentation in single laparoscopic frames. For the SAM-based benchmark, prompts were generated from the original ground-truth masks to ensure controlled comparison and isolate prompt sensitivity. We tested point, box, box+point, and box+positive/negative point prompts, where supported (Table 1). Positive points were selected using a distance transform as the foreground pixel farthest from the background, improving stability for irregular masks. Boxes tightly enclosed the ground-truth masks. For box+positive/negative point prompting, boxes were expanded by 0.5 relative padding around the ground-truth masks, and four negative points were sampled from the surrounding top, bottom, left, and right background regions. This setup quantified prompt sensitivity and identified the most reliable prompt type.

Table 1. Prompt types supported by the evaluated SAM-based models. Neg.: negative points. × denotes unsupported or not standard prompt format.

Model	Point	Box	Box+Point	Box+Point+ Neg.
SAM2	√	√	√	√
MedSAM	×	√	×	×
SAM-Med2D	√	√	×	×
SurgiSAM2	√	√	√	√

Second, we trained the fully automatic supervised models (CNN-based, Transformer-based, self-configuring biomedical, and YOLO-based) and evaluated them on the same held-out test sets. All models were trained or evaluated as binary lesion segmentation models. For supervised models, one model was trained per dataset split, and the checkpoint with the best validation Dice score was retained for testing. Early stopping was applied with a patience of 20 epochs where applicable. For nnU-Net v2, the default 2D configuration was trained for a maximum of 100 epochs to maintain a comparable computational budget across supervised models; the best validation checkpoint was retained for testing. Training and validation losses were logged for all trained models.

Third, we performed a hybrid automatic-refinement experiment using the best supervised model to generate lesion candidate masks. Tight boxes around these candidates were used to prompt the best performing SAM-based model (AutoBox), and the resulting masks were evaluated as automatic refinements of the supervised predictions. We also tested fallback acceptance, retaining the SAM-refined mask only when it agreed with the supervised candidate (Dice $\geq$ 0.80); otherwise, the original supervised prediction was kept (Fallback). This tested whether manually prompted SAM performance transfers to fully automatic lesion segmentation and mask refinement. The same was tested by adding positive/negative points to the derived box prompt from the initial candidate mask (PosNeg Fallback).

2.4 System Configuration

All experiments were conducted on a machine equipped with an NVIDIA GeForce RTX 5090 GPU, 32GB GDDR7 memory, running Python 3.11, PyTorch 2.11.0, and CUDA 12.8. Experiments were executed in model-specific virtual environments. All scripts, logs, configurations and model checkpoints are documented under the project repository to ensure reproducibility (see *GitHub*).

2.5 Evaluation Metrics and Statistical Testing

Segmentation performance was evaluated at the image level using Dice similarity coefficient, intersection-over-union (IoU), precision, and recall. Dice and IoU measured spatial overlap, recall quantified lesion sensitivity, and precision captured false-positive behavior. Metrics were computed separately for each supported SAM prompt setting and consistently applied to supervised and hybrid outputs. Results are reported as mean ± standard deviation on validation and held-out test sets, with test results used for final comparison.

We also recorded training/validation losses, completed epochs, early stopping, selected checkpoints, inference times, per-image metrics, predicted masks, visualizations, and summary spreadsheets. Inference time was measured for end-to-end mask generation, including candidate prediction and SAM refinement in the hybrid experiments (see *GitHub*).

Statistical analyses used paired image-level metrics, since all methods were evaluated on the same image sets. Prompt comparisons included box versus point, box+point versus box, and box+positive/negative versus box or box+point, for compatible models. Two-method comparisons used two-sided Wilcoxon signed-rank tests; comparisons across more than two related methods used Friedman tests followed by post-hoc Wilcoxon tests when significant. Hybrid variants were compared with each other and against SegFormer on held-out test images. Multiple comparisons were corrected using Benjamini–Hochberg or Holm–Bonferroni adjustment, with significance defined as $p < 0.05$.

3 Results

Across compatible SAM-based models, box prompts consistently outperformed point-only prompts. On the test sets, box prompting increased mean Dice by 0.179 on ENID and 0.410 on GLENDA-clean ($p_{\text{adj}} = 6.75 \times 10^{-32}$), with similar significant gains in mean IoU and recall. Adding a positive point to the box produced negligible Dice changes (< 0.005), while box+positive+negative prompting significantly reduced Dice on ENID ($\Delta = -0.006$, $p_{\text{adj}} = 0.024$) and GLENDA-clean ($\Delta = -0.033$, $p_{\text{adj}} = 3.16 \times 10^{-7}$), accompanied by recall reductions. These results indicate that boxes provide the most reliable lesion localization, whereas negative points can over-constrain masks and increase false negatives. Box prompting was therefore selected as the main SAM prompt strategy for subsequent comparisons.

Comparing the SAM-based models using box-prompts, SAM2 and SurgiSAM2 achieved the highest Dice scores, with SAM2 performing best on ENID, achieving significantly higher Dice and IoU than SurgiSAM2 ($p_{\text{adj}} = 0.024$ and $p_{\text{adj}} = 0.026$, respectively). On the more heterogeneous GLENDA-clean set, SurgiSAM2 performed best, with significantly higher recall than SAM2 ($p_{\text{adj}} = 0.001$) and significantly higher Dice, IoU, and recall than the medical SAM variants ($p_{\text{adj}} \leq 2.95 \times 10^{-5}$)(Table 2).

Table 2. Model size and performance on ENID and GLENDA-clean test-sets, reported as mean Dice, IoU, recall, and inference time (± standard deviation. Hybrid inference time includes supervised prediction, automatic prompt generation, and SAM-based refinement). Best Dice scores per dataset are shown in **bold**.

Model	Params	ENID				GLENDA-clean			
	(M)	Dice	IoU	Rec.	Time (s)	Dice	IoU	Rec.	Time (s)
SAM2	80.80	**0.87**	0.77	0.87	0.012 ± 0.007	**0.75**	0.61	0.70	0.011 ± 0.006
MedSAM	93.70	0.79	0.67	0.69	0.006 ± 0.002	0.69	0.54	0.59	0.006 ± 0.002
SAM-Med2D	≈94.00	0.78	0.65	0.79	0.003 ± 0.000	0.69	0.54	0.62	0.003 ± 0.001
SurgiSAM2	80.80	**0.85**	0.75	0.85	0.006 ± 0.002	**0.77**	0.64	0.76	0.006 ± 0.002
DeepLabV3+	≈39.60	0.73	0.61	0.76	0.003 ± 0.000	0.40	0.27	0.53	0.002 ± 0.000
SegFormer	3.80	**0.82**	0.71	0.84	0.005 ± 0.000	**0.47**	0.34	0.54	0.004 ± 0.000
YOLO11s-seg	10.10	0.66	0.55	0.75	0.011 ± 0.001	0.17	0.11	0.17	0.010 ± 0.001
U-Net++	≈26.10	**0.81**	0.70	0.83	0.005 ± 0.000	**0.39**	0.27	0.42	0.005 ± 0.000
nnU-Net v2 2D	≈31.00	0.73	0.62	0.73	7.965 ± 0.176	0.27	0.19	0.30	7.974 ± 0.435
Hybrid	≈84.60	0.78	0.66	0.78	0.052 ± 0.031	0.43	0.30	0.48	0.053 ± 0.017
Hybrid + Fallback	≈84.60	0.80	0.68	0.79	0.071 ± 0.042	0.47	0.34	0.53	0.061 ± 0.025
Hybrid +PosNeg. +Fallback	≈84.60	0.79	0.67	0.79	0.063 ± 0.042	0.47	0.34	0.53	0.062 ± 0.022

Among trained models, SegFormer showed the strongest overall performance across both datasets, with higher Dice, IoU, and recall while maintaining low inference time, comparable to U-Net++ on ENID; but significantly outperformed nnU-Net v2 in Dice, IoU, and recall ($p_{\text{adj}} = 0.026$, $p_{\text{adj}} = 0.023$, and $p_{\text{adj}} = 0.0145$ respectively). On GLENDA-clean, SegFormer achieved the best trained-model Dice and IoU, significantly exceeding all other trained baselines for both overlap metrics ($p_{\text{adj}} \leq 0.003$) (Table 2). Qualitative examples confirmed these trends, with representative SAM-based and trained-model predictions shown in Figures 1 and 2.

Hybrid pipelines were evaluated to test whether SurgiSAM2 refinement could improve SegFormer predictions and whether fallback acceptance or additional negative-point prompts provided further benefit. Direct refinement (AutoBox) degraded performance relative to SegFormer, reducing Dice by 0.034 on ENID and 0.040 on GLENDA-clean ($p_{\text{adj}} \leq 0.005$), with similar significant decreases in IoU and recall.

Fallback acceptance improved the hybrid over direct AutoBox refinement, increasing Dice by 0.018 on ENID ($p_{\text{adj}} = 0.004$) and 0.034 on GLENDA-clean ($p_{\text{adj}} = 0.008$), with corresponding IoU gains of 0.023 and 0.031. However, com-

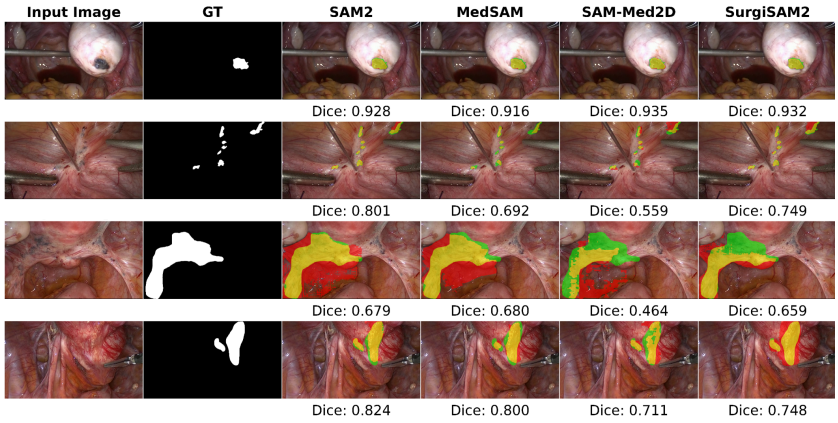

Fig. 1. Qualitative comparison of selected SAM-based models on challenging unseen samples from ENID (top two rows) and GLENDA-clean (bottom two rows), representing single and multifocal lesions, irregular lesions resembling surrounding tissue, and dark regions outside the annotated ground truth. Overlays show ground truth in green, predictions in red, and overlap in yellow.

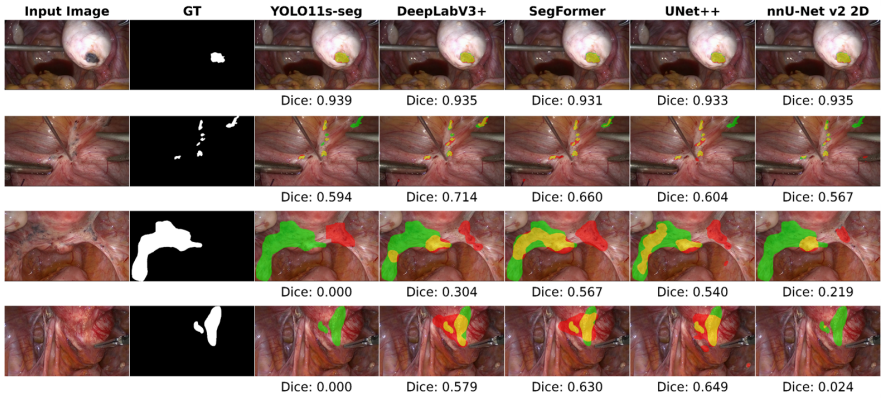

Fig. 2. Qualitative comparison of selected trained models on challenging unseen samples from ENID (top two rows) and GLENDA-clean (bottom two rows). Overlays show ground truth in green, predictions in red, and overlap in yellow.

pared with SegFormer alone, the fallback hybrid remained worse on ENID, including a significant Dice decrease ($\Delta = -0.016$, $p_{\text{adj}} = 0.0009$), and showed a small recall decrease on GLENDA-clean ($\Delta = -0.012$). The PosNeg fallback variant did not significantly improve over box-only fallback. Overall, fallback reduced the degradation from SAM refinement, but no hybrid variant significantly outperformed SegFormer in Dice or IoU (Table 2 and Figure 3).

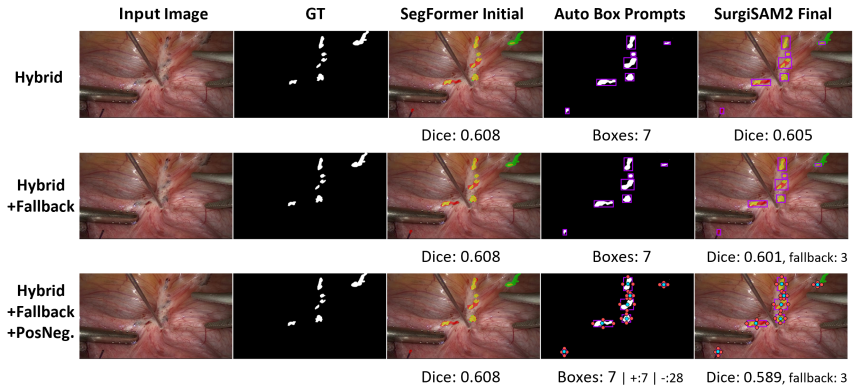

Fig. 3. Qualitative comparison of hybrid refinement strategies on an ENID test image using SegFormer predictions as automatic prompts for SurgiSAM2. Rows show direct SegFormer+SurgiSAM2 refinement, fallback refinement, and fallback with positive/negative points. Values below each output report Dice before/after refinement, number of fallbacks to SegFormer, and number of positive (+) and negative (-) points. Overlays show ground truth in green, predictions in red, and overlap in yellow.

4 Discussion and Conclusion

This study shows that prompt design strongly affects SAM-based laparoscopic endometriosis segmentation. Box prompts consistently outperformed point, while adding positive or negative points did not improve performance. Negative background points reduced Dice and recall, suggesting they over-constrained predictions and removed true lesion pixels. This is plausible for endometriosis lesions, which are often small, irregular, weakly contrasted, and visually ambiguous against surrounding tissue, peritoneum, blood, fibrosis, shadows, or specular highlights.

Performance was strongly dataset-dependent. SAM2 performed best on ENID, whereas SurgiSAM2 was more robust on the heterogeneous GLENDA-clean set, suggesting that surgical-domain adaptation may help under variable laparoscopic appearance. Among fully automatic models, SegFormer achieved the strongest overall performance, particularly on GLENDA-clean. Its transformer-based encoder may better capture global context and heterogeneous lesion appearance than purely convolutional baselines, supporting task-specific supervised learning as the most reliable strategy for deployable automatic segmentation in this benchmark.

The hybrid experiments showed that SegFormer predictions can automatically prompt SurgiSAM2, but refinement did not outperform SegFormer alone. Performance was limited by candidate quality: lesions missed by SegFormer could not be recovered, while inaccurate or fragmented candidates generated suboptimal prompts and propagated errors. SurgiSAM2 also sometimes reduced lesion coverage by shrinking otherwise adequate masks. Fallback acceptance limited

this degradation and recovered performance closer to the supervised baseline, suggesting that hybrid prompting may still be useful in annotation-limited or semi-automatic workflows, but likely requires selective or uncertainty-aware refinement or learned prompt generation, rather than unconditional SAM post-processing.

This study is limited by small datasets, annotation variability, image-quality differences, and heterogeneous lesion appearance. SAM-based models were intentionally evaluated without fine-tuning to assess out-of-the-box promptability, although task-specific adaptation may improve performance. The hybrid pipeline was further limited by automatic prompt quality, since missed lesions could not be recovered and inaccurate candidate boxes could propagate errors. Future work should investigate fine-tuned surgical foundation models, robustness to domain shift, learned or uncertainty-aware prompting, temporally consistent video segmentation, and validation on larger harmonized laparoscopic datasets.

Acknowledgments. This study was supported by the Bavarian State Ministry of Health, Care and Prevention (StMGP) in the context of the project EndoKI.

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

References

1. Bondarenko, A., Jumutc, V., Netter, A., Duchateau, F., Abrão, H.M., Noorzadeh, S., Giacomello, G., Ferrari, F., Bourdel, N., Kirk, U.B., Bļizņuks, D.: Object detection in laparoscopic surgery: A comparative study of deep learning models on a custom endometriosis dataset. Diagnostics **15**(10), 1254 (2025)
2. Chen, H., Gou, L., Fang, Z., Dou, Q., Chen, H., Chen, C., Qiu, Y., Zhang, J., Ning, C., Hu, Y., Deng, H., Yu, J., Li, G.: Artificial intelligence assisted real-time recognition of intra-abdominal metastasis during laparoscopic gastric cancer surgery. npj Digital Medicine **8**(1), 9 (2025)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision. pp. 833–851. Springer (2018)
4. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: SAM-Med2D (2023), arXiv preprint arXiv:2308.16184
5. Figueredo, W.K.R., Silva, A.C., de Paiva, A.C., Diniz, J.O.B., Brandão, A., Oliveira, M.A.P.: Automatic segmentation of deep endometriosis in the rectosigmoid using deep learning. Image and Vision Computing **151**, 105261 (2024)
6. Gül, S., Cetinel, G., Aydin, B.M., Akgün, D., Öztaş Kara, R.: YOLOSAMIC: A hybrid approach to skin cancer segmentation with the segment anything model and YOLOv8. Diagnostics **15**(4), 479 (2025)
7. Isensee, F., Jäger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
8. Jocher, G., Qiu, J.: Ultralytics YOLO11. `https://github.com/ultralytics/ultralytics` (2024)

9. Kamtam, D.N., Shrager, J.B., Malla, S.D., Wang, X., Lin, N., Cardona, J.J., Yeung-Levy, S., Hu, C.: A fine-tuned foundational model SurgiSAM2 for surgical video anatomy segmentation and detection. Scientific Reports **15**(1), 35961 (2025)
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
11. Kolbinger, F.R., Rinner, F.M., Jenke, A.C., Carstens, M., Krell, S., Leger, S., Distler, M., Weitz, J., Speidel, S., Bodenstedt, S.: Anatomy segmentation in laparoscopic surgery: Comparison of machine learning and human expertise–an experimental study. International Journal of Surgery **109**(10), 2962–2974 (2023)
12. Leibetseder, A., Kletz, S., Schoeffmann, K., Keckstein, S., Keckstein, J.: GLENDA: Gynecologic laparoscopy endometriosis dataset. In: MultiMedia Modeling. pp. 439–450. Springer (2020)
13. Leibetseder, A., Schoeffmann, K., Keckstein, J., Keckstein, S.: Endometriosis detection and localization in laparoscopic gynecology. Multimedia Tools and Applications **81**(5), 6191–6215 (2022)
14. Li, H., Zhang, D., Yao, J., Han, L., Li, Z., Han, J.: ASPS: Augmented segment anything model for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 118–128. Springer (2024)
15. Lin, X., Xiang, Y., Yu, L., Yan, Z.: Beyond adapting SAM: Towards end-to-end ultrasound image segmentation via auto prompting. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 24–34. Springer (2024)
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
17. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: An experimental study. Medical Image Analysis **89**, 102918 (2023)
18. Nasirihaghighi, S., Ghamsarian, N., Peschek, L.S., Munari, M., Husslein, H., Sznitman, R., Schoeffmann, K.: GynSurg: A comprehensive gynecology laparoscopic surgery dataset. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 13141–13147. ACM (2025)
19. Netter, A., Noorzadeh, S., Duchateau, F., Abrao, H., Desternes, J., Peyras, J., Pouly, J.L., Abrão, M.S., Bokor, A., Kirk, U.B., Agostini, A., Courbiere, B., Canis, M., Bartoli, A., Bourdel, N., Group, F.P.W.: Initial results in the automatic visual recognition of endometriosis lesions by artificial intelligence during laparoscopy: A proof-of-concept study. Journal of Minimally Invasive Gynecology **32**(12), 1118–1125 (2025)
20. Podda, A.S., Balia, R., Barra, S., Carta, S., Neri, M., Guerriero, S., Piano, L.: Multi-scale deep learning ensemble for segmentation of endometriotic lesions. Neural Computing and Applications **36**(24), 14895–14908 (2024)
21. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: International Conference on Learning Representations (2025)
22. Sheng, Y., Bano, S., Clarkson, M.J., Islam, M.: Surgical-DeSAM: Decoupling SAM for instrument segmentation in robotic surgery. International Journal of Computer Assisted Radiology and Surgery **19**(7), 1267–1271 (2024)

23. Wang, C.Y., Yeh, I.H., Liao, H.Y.M.: YOLOv9: Learning what you want to learn using programmable gradient information. In: European Conference on Computer Vision. pp. 1–21. Springer (2024)
24. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems. vol. 34, pp. 12077–12090 (2021)
25. Yuan, C., Jiang, J., Yang, K., Wu, L., Wang, R., Meng, Z., Ping, H., Xu, Z., Zhou, Y., Song, W., Wang, H., Jin, Y., Dou, Q., Ban, Y.: Systematic evaluation and guidelines for segment anything model in surgical video analysis. npj Digital Surgery **1**(1), 2 (2026)
26. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: U-Net++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 3–11. Springer (2018)