

Surgical Anatomy Recognition with Context Learning using Foundation Representations

Ronald L. P. D. de Jong^{1*}, Tim J. M. Jaspers^{2*}, Raf A. H. Vervoort², Aron F. H. A. Bakker^{4,5}, Yiping Li¹, Jip L. Tolenaar³, Jelle P. Ruurda³, Willem M. Brinkman⁴, Josien P. W. Pluim¹, Marcel Breeuwer¹, Daan de Geus⁶, and Fons van der Sommen²

¹ Department of Biomedical Engineering, Medical Image Analysis, Eindhoven University of Technology, Eindhoven, The Netherlands

r.l.p.d.d.jong@tue.nl

² Department of Electrical Engineering, Architectures for Reliable Image Analysis, Eindhoven University of Technology, Eindhoven, The Netherlands

t.j.m.jaspers@tue.nl

³ Department of Surgery, University Medical Center Utrecht, Utrecht, The Netherlands

⁴ Department of Oncological Urology, University Medical Center Utrecht, Utrecht, The Netherlands

⁵ Department of Urology, Catharina Hospital, Eindhoven, The Netherlands

⁶ Department of Electrical Engineering, Mobile Perception Systems Lab, Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract. Accurate recognition of anatomical structures is essential for safe and effective minimally invasive surgery (MIS), yet it remains under-explored in surgical computer vision due to limited annotated data and methods tailored primarily to natural scenes. In this work, we present a combined dataset and model framework to advance anatomy-aware perception in MIS. First, we introduce ATLAS-120k, a large-scale clip-level semantic segmentation dataset comprising over 120,000 annotated frames from 100 surgical videos spanning 14 procedures and multiple modalities, including laparoscopic and robot-assisted surgery. The dataset captures substantial procedural variability and was created using a scalable annotation pipeline that integrates expert manual labeling, automated propagation, iterative refinement, and surgeon verification to ensure high-quality annotations. Second, we propose ATLAS (Anatomy Recognition with Context Learning using Foundation Representations), a video semantic segmentation model specifically designed for surgical anatomy recognition. Unlike conventional approaches that emphasize object tracking, ATLAS leverages foundation-model embeddings together with lightweight temporal reasoning to incorporate contextual cues such as procedure type, surgical phase, and short-term visual memory. This design enables temporally consistent and accurate predictions while maintaining real-time feasibility. Together, the dataset and model establish a practical foundation for robust surgical scene understand-

* R. L. P. D. de Jong and T. J. M. Jaspers contributed equally to this work.

ing and support the development of clinically applicable guidance systems for minimally invasive surgery. The models, dataset annotations and annotation platform are publicly available at: <https://github.com/TimJaspers0801/ATLAS>.

Keywords: Video semantic segmentation · Foundation models · Surgical computer vision.

1 Introduction

In recent years the field of surgical computer vision has advanced rapidly, driven by breakthroughs in deep learning and image analysis. These methods promise tangible benefits for minimally invasive surgery (MIS), from real-time guidance to improved precision and patient safety, yet clinical translation remains limited. A key bottleneck is the lack of large, diverse, and richly annotated datasets that capture the variability of real surgical practice and thus enable robust, generalizable models. This scarcity has constrained progress despite growing interest in anatomy-aware guidance systems [3, 12, 15].

Most prior work has concentrated on well-defined tasks such as surgical phase recognition [13, 21], tool segmentation [1], and instrument tracking [25]. While these tasks are important building blocks for computer-assisted surgery, accurate recognition of anatomical structures is both a distinct and underexplored challenge: in MIS, surgeons must reason about anatomy with limited viewpoints and without tactile feedback, so robust anatomical perception is essential for safe navigation. Existing anatomy datasets have begun to address this need, but many suffer from limited procedural diversity, relatively few videos or frames, and incomplete coverage of anatomical classes, factors that limit model generalization in realistic settings [4, 7, 14].

To address these gaps we contribute two complementary advances. First, we introduce ATLAS-120k, a large clip-level surgical anatomy segmentation dataset that substantially broadens procedural and technological diversity: it comprises annotations drawn from 100 videos spanning 14 distinct procedures, and includes both laparoscopic and robot-assisted MIS, resulting in over 120k annotated frames.

Second, we present ATLAS (Anatomy Recognition with Context Learning using Foundation Representations), a video semantic segmentation model designed specifically for surgical anatomy recognition. Surgical anatomy segmentation, unlike video segmentation for most natural scenes, depends heavily on contextual cues that surgeons use in practice. Specifically, the type of procedure and surgical phase (i.e., a short, distinct stage of the surgery) are critical for understanding which anatomical categories may appear. ATLAS integrates lightweight temporal and tracking components to capture this procedure- and phase-dependent context and leverages strong embeddings from a surgical foundation model [8] to further enhance its domain knowledge. The resulting architecture enables practical real-time operation, while substantially improving anatomical consistency and accuracy across clips and procedures.

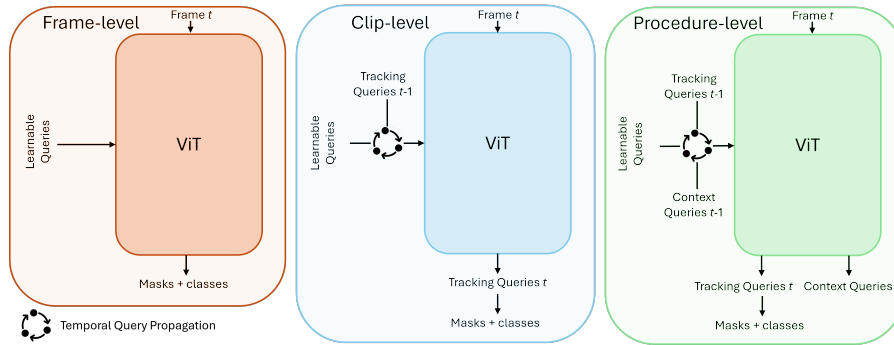


Fig. 1: Left: the frame-level semantic segmentation model. Middle: the clip-level model with integrated tracking queries. Right: the procedure-level model, incorporating context queries.

Together, ATLAS-120k and the ATLAS model provide a paired dataset and method package designed to advance anatomy-aware surgical perception: the dataset expands the scope and realism of available supervision, and the model demonstrates how foundation-level representations plus compact temporal reasoning can deliver accurate, real-time anatomical segmentation across a wide range of MIS procedures. These contributions aim to close the gap between research prototypes and clinically useful, anatomy-aware guidance systems and represent a step toward safer, more intelligent assistance in the operating room.

2 Methods

Overview. Video semantic segmentation in surgical settings presents unique challenges compared to natural videos. While tracking objects across frames is important, successful surgical segmentation also requires domain-specific knowledge: surgeons rely on procedural understanding and awareness of the current surgical phase to identify anatomical structures accurately. Capturing this contextual information is therefore critical.

We build on EoMT [11], a state-of-the-art image segmentation model that leverages strong pretrained visual embeddings. EoMT uses a ViT encoder to process learnable queries, each representing a single object, and predicts a class and segmentation mask for each of these queries—conducting dense segmentation without complex decoders. VidEoMT [17] extends EoMT to videos by propagating queries from frame $t - 1$ to frame t , allowing segmentation and tracking over time. We further build on VidEoMT by introducing context queries that augment the segmentation and tracking queries. These context queries encode global procedural and temporal information, enabling the model to incorporate procedural context, which VidEoMT does not capture, for more accurate and consistent segmentation. Figure 1 provides an overview of the proposed method.

Transformer Input. For each video frame t , the input to the transformer consists of

$$X_t = (P_t, Q_0^{\text{seg}}, Q_0^{\text{proc}}, Q_0^{\text{phase}}, \hat{Q}_{t-1}^{\text{seg}}, \hat{Q}_{t-1}^{\text{proc}}, \hat{Q}_{t-1}^{\text{phase}}), \quad (1)$$

where P_t are visual patch tokens, Q_0^{seg} are learnable segmentation queries, and $\hat{Q}_{t-1}^{\text{seg}}$ are propagated segmentation queries from the previous frame. We additionally introduce learnable initial context queries Q_0^{proc} and Q_0^{phase} that are refined over time. Previous-frame context queries ($t-1$) provide temporal continuity. All queries are processed jointly by a single transformer encoder. Because attention is global, segmentation queries can directly attend to both visual evidence and contextual queries, enabling context-aware predictions.

Procedure Context Queries. Procedure queries encode high-level surgical context that remains relatively stable across a video. For each time step, the updated procedure query embeddings $\hat{Q}_t^{\text{proc}}$ are passed through a multilayer perceptron (MLP) to predict a procedure label

$$\hat{y}_t^{\text{proc}} = \text{MLP}(\hat{Q}_t^{\text{proc}}). \quad (2)$$

We compute the cross-entropy loss between $\hat{y}_t^{\text{proc}}$ and ground-truth label y_t^{proc} :

$$\mathcal{L}_{\text{proc}} = \text{CE}(\hat{y}_t^{\text{proc}}, y_t^{\text{proc}}). \quad (3)$$

This auxiliary supervision encourages the procedure queries to represent global scene context useful for segmentation, such as instrument presence or anatomical exposure patterns.

Phase Context Queries. Phase queries capture finer temporal variations. For each clip, we obtain a phase embedding

$$z_t = \text{MLP}(\hat{Q}_t^{\text{phase}}). \quad (4)$$

We apply an InfoNCE loss that induces an attractive force between embeddings from temporally adjacent frames within the same clip (z_{t+}), while repelling embeddings from different clips (z_j) or distant timestamps using scaling factor τ :

$$\mathcal{L}_{\text{phase}} = -\log \frac{\exp(z_t \cdot z_{t+}/\tau)}{\sum_j \exp(z_t \cdot z_j/\tau)}. \quad (5)$$

This self-supervised signal encourages phase queries to encode temporally coherent surgical state information.

Temporal Query Propagation. To ensure short-term memory and temporal consistency, queries are propagated over time in a manner similar to VidEoMT [17]. The input queries Q_t for the encoder-only model at frame t are a function of the output queries $\hat{Q}_{t-1}$ of the model at frame $t-1$ and the initial learnable queries Q_0 :

$$Q_t = \text{Linear}(\hat{Q}_{t-1}) + Q_0, \quad (6)$$

where Linear is a linear layer and

$$\hat{Q}_{t-1} = [\hat{Q}_{t-1}^{\text{seg}} \hat{Q}_{t-1}^{\text{proc}} \hat{Q}_{t-1}^{\text{phase}}], \text{ and } Q_0 = [Q_0^{\text{seg}} Q_0^{\text{proc}} Q_0^{\text{phase}}]. \quad (7)$$

Loss function. The loss function follows the original EoMT [11] loss, with additional procedure and phase losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{bce}} \mathcal{L}_{\text{bce}} + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}} + \lambda_{\text{ce}} \mathcal{L}_{\text{ce}} + \lambda_{\text{proc}} \mathcal{L}_{\text{proc}} + \lambda_{\text{phase}} \mathcal{L}_{\text{phase}}, \quad (8)$$

where the weights λ_{bce} , λ_{dice} , λ_{ce} , λ_{proc} , and λ_{phase} were set to 5.0, 5.0, 2.0, 1.0, and 0.1, respectively.

Pretraining. To improve model performance, the model was initialized with in-domain pretrained weights following the method described in [9].

3 Anatomy Segmentation Dataset

Figure 2 shows representative examples from the ATLAS-120k dataset, with at least one example per procedure. Table 1 compares ATLAS-120k with existing surgical datasets. ATLAS-120k distinguishes itself by including 14 distinct surgical procedures, a substantial expansion over prior datasets, which typically cover only a single procedure. Moreover, ATLAS-120k integrates both laparoscopic and robot-assisted minimally invasive surgical videos, thereby capturing a broader spectrum of procedural and technological variability. The Dresden Surgical Anatomy Dataset (DSAD) [4] also contains a large number of frames annotated with multiple anatomical classes; however, most DSAD frames depict isolated anatomical segmentations rather than comprehensive semantic segmentation of the full surgical scene. Both the Endoscopes-Seg50 [14] and Cholec-Seg8k [7] datasets focus exclusively on minimally invasive surgery and contain fewer annotated frames than ATLAS-120k.

A total of 100 videos were sourced from the GSViT dataset [19], from which 14 distinct procedures were selected for annotation. Rather than annotating full videos, clips of varying lengths were extracted. To standardize the dataset, all videos were downsampled to 15 fps, corresponding to the lowest original frame rate observed. Spatial resolutions varied from 480×640 to 1080×1920 pixels. Figure 3a presents the distribution of clip durations, showing that most clips

Table 1: Comparison of surgical video segmentation dataset characteristics. MIS indicates minimally invasive surgery; RA indicates robot-assisted procedures.

Dataset	Included procedures	# classes	# videos	# clips	# frames	MIS	RA
Endoscopes-Seg50 [14]	1	6	50	-	493	✓	
CholecSeg8k [7]	1	12	17	-	8,080	✓	
DSAD [4]	1	11	32	-	14,625		✓
ATLAS-120k	14	42	100	502	121,018	✓	✓



Fig. 2: Examples of randomly selected frames and annotations, with at least one example from each included procedure and class.

contain fewer than 400 frames (<30 seconds). Figure 3b illustrates the number of clips and frames per procedure, with cholecystectomy contributing the largest share. Annotations were performed by three surgical research fellows under the supervision of three experienced surgeons (each with more than 10 years of experience). Although the primary objective was anatomical structure recognition, all visible surgical instruments were also annotated, enabling additional applications such as surgical tool segmentation. The annotation workflow consisted of two phases. First, the initial frame of each clip was manually annotated with high precision. Second, a lightweight object-tracking model [6] was used to propagate

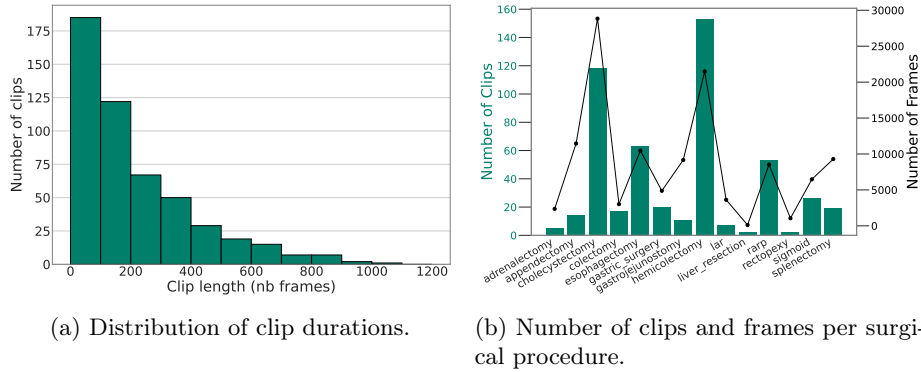


Fig. 3: ATLAS-120k dataset characteristics.

these annotations across subsequent frames. Any propagation errors were manually corrected to ensure high annotation quality. To further improve tracking performance, the model was iteratively fine-tuned on the annotated data after approximately 10k, 25k, and 50k frames, using the original training parameters described in [6]. Finally, the first frame of every clip was independently reviewed by at least one experienced surgeon to verify the correctness and completeness of the anatomical labels. The annotations of the dataset will be released as open source under a CC-BY-NC-SA-4.0 license to facilitate research in this direction.

4 Experiments & Results

Dataset. The ATLAS-120k dataset was partitioned into patient-level training (70 videos), validation (10 videos), and test (20 videos) splits, with a balanced distribution of procedures across the splits. To ensure consistency, we restricted model training to 30 classes by excluding categories that were not represented in all subsets and consolidating semantically similar classes.

Implementation details. All experiments were conducted on an NVIDIA H100 GPU. The model was trained for 10 epochs with a batch size of 24 and a clip length of 5 frames. The architecture incorporated both procedural and temporal context representations, using one procedure context query and four phase context queries. In addition, 100 segmentation queries were employed. All hyperparameters were kept fixed across experiments unless otherwise stated. Data leakage was avoided by excluding ATLAS-120k from pretraining.

Evaluation. All models are evaluated using detection and segmentation metrics. Detection performance is measured with COCO-style mean Average Precision (AP), including AP, AP75, and AP50, while segmentation quality is assessed using the Dice coefficient. Temporal stability is evaluated with the mean Video Consistency metric (mVC) [16], computed over sliding windows of 12 and 24 consecutive frames. mVC quantifies the proportion of pixels whose predicted labels remain constant across a window, restricted to regions where ground-truth labels are temporally consistent and non-background. The final mVC score is obtained by averaging consistency across all valid windows and clips.

State-of-the-art comparison. Table 2 compares ATLAS to state-of-the-art methods. Surgical foundation models such as SurgeNetXL [8], LEMON [5], and GSViT [19] were excluded from this comparison due to overlap between their pretraining data and the ATLAS-120k dataset. ATLAS consistently outperforms both natural-image and in-domain surgical foundation models across detection and segmentation metrics. While DINO variants [18, 20] show limited transfer to endoscopic videos, in-domain models such as EndoViT [2] and SurgeNet improve performance but fall short of the strongest ATLAS variants. Our ViT-L model achieves 0.64 AP, 0.49 mDice, and 0.79 mVC24, showing that combining foundation embeddings with temporal and procedural queries yields strong gains. Combined with an inference speed of 64 FPS (benchmarked on an NVIDIA H100), this enables real-time performance. Although the absolute mDice may appear moderate, it reflects the challenging nature of ATLAS-120k,

Table 2: Quantitative comparison of foundation models for segmentation. Models are grouped by pretraining domain: natural-image foundation models (top), in-domain pretrained medical models (middle), and our proposed ATLAS models (bottom). For DINOv2 and DINOv3, we train a linear layer attached to the ViT.

Model	Backbone	#PARAMS (M) ↓	AP ↑	AP75 ↑	AP50 ↑	mDice ↑	mVC12 ↑	mVC24 ↑
DINOv2 [18]	ViT-B	87	0.32	0.29	0.61	0.09	0.69	0.57
DINOv2 [18]	ViT-L	304	0.29	0.25	0.61	0.10	0.61	0.49
DINOv3 [20]	ViT-B	86	0.41	0.43	0.77	0.27	0.72	0.63
DINOv3 [20]	ViT-L	303	0.43	0.44	0.74	0.28	0.70	0.60
SAM2-UNet [24]	Hiera-L	217	0.45	0.46	0.76	0.24	0.63	0.52
SAM3-UNet [23]	Hiera-L	449	0.48	0.51	0.76	0.35	0.73	0.64
EndoFM [22]	ViT-B	141	0.24	0.17	0.55	0.11	0.52	0.39
GastroNet5M [10]	ViT-B	86	0.36	0.33	0.66	0.18	0.66	0.54
EndoViT [2]	ViT-B	112	0.46	0.49	0.79	0.32	0.73	0.64
SurgeNet-pvt [8]	PVTv2	27	0.47	0.49	0.79	0.32	0.78	0.70
SurgeNet-conv [8]	ConvNeXtv2	30	0.50	0.52	0.78	0.31	0.66	0.54
SurgeNet-ca [8]	CaFormer-S18	26	0.51	0.55	0.80	0.32	0.61	0.49
ATLAS (ours)	ViT-S	24	0.52	0.54	0.80	0.31	0.75	0.66
ATLAS (ours)	ViT-B	94	0.60	0.67	0.85	0.42	0.81	0.74
ATLAS (ours)	ViT-L	318	0.64	0.70	0.89	0.49	0.85	0.79

which comprises 30 classes with highly imbalanced and low-prevalence anatomical structures. This class diversity and long-tail distribution make ATLAS-120k a realistic and demanding benchmark for future anatomy segmentation models.

Ablation study. Table 3 quantifies the contribution of each component in ATLAS. Starting from the EoMT baseline with default DINOv3 weights, in-domain pretraining provides a substantial boost in both detection and segmentation metrics. Adding temporal query propagation further improves detection performance, as reflected by higher AP values. However, it reduces overlap and temporal consistency metrics, which may be due to boundary drift in propagated frames. Incorporating context queries restores and surpasses all metrics, improving AP by over 20% relative to the baseline while increasing the parameter count by less than 1%. These results highlight that both temporal modeling and procedural context are critical for robust surgical video segmentation, a capability enabled by the scale and procedural diversity of ATLAS-120k. Additionally, Table 3 shows the benefits of advanced pretraining methods.

Table 3: Ablation studies showing: (1) the effect of tracking and context queries on surgical video segmentation, (2) the benefits of advanced pretraining.

Method	#PARAMS (M) ↓	AP ↑	AP75 ↑	AP50 ↑	mDice ↑	mVC12 ↑	mVC24 ↑
EoMT ViT-L (Baseline)	315	0.53	0.58	0.81	0.35	0.76	0.68
→ w/In-domain pretraining	315	0.57	0.62	0.85	0.43	0.80	0.74
→ w/Temporal query propagation	315	0.59	0.63	0.86	0.36	0.78	0.71
→ w/Context queries (ATLAS)	318	0.64	0.70	0.89	0.49	0.85	0.79
ATLAS ViT-B (DINOv1)	94	0.52	0.56	0.80	0.30	0.82	0.75
→ w/DINOv2	92	0.55	0.59	0.81	0.45	0.85	0.79
→ w/DINOv3	94	0.60	0.67	0.85	0.42	0.81	0.74

5 Conclusion

We presented ATLAS-120k, a large-scale, diverse surgical video dataset, and ATLAS, a context-aware video segmentation model for anatomy recognition in minimally invasive surgery. By combining foundation-model embeddings with temporal and procedural context queries, ATLAS delivers accurate segmentation in real time. Experiments show that both temporal propagation and context modeling are essential for robust anatomy understanding. Together, the dataset and model provide a foundation for future research in anatomy-aware surgical perception and clinically relevant guidance systems.

Acknowledgments. This study was funded by Stichting Hanarth Fonds (2022-13) and the Dutch Research Council (KICH1.ST03.21.019). Furthermore, we thank SURF (www.surf.nl) for their support in using the national supercomputer Snellius.

Disclosure of Interests. The authors have no competing interests.

References

1. Allan, M., Shvets, A., Kurmann, T., et al.: 2017 robotic instrument segmentation challenge (2019), <https://arxiv.org/abs/1902.06426>
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). <https://doi.org/10.1007/s11548-024-03091-5>, <https://doi.org/10.1007/s11548-024-03091-5>
3. den Boer, R.B., Jaspers, T.J.M., de Jongh, C., et al.: Deep learning-based recognition of key anatomical structures during robot-assisted minimally invasive esophagectomy. *Surgical Endoscopy* **37**(7), 5164–5175 (July 2023). <https://doi.org/10.1007/s00464-023-09990-z>
4. Carstens, M., Rinner, F.M., Bodenstedt, S., et al.: The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data* **10**(1), 3 (2023). <https://doi.org/10.1038/s41597-022-01719-2>
5. Che, C., Wang, C., Vercauteren, T., Tsoka, S., Garcia-Peraza-Herrera, L.C.: Lemon: A large endoscopic monocular dataset and foundation model for perception in surgical settings. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 42659–42669 (June 2026)
6. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3151–3161 (June 2024)
7. 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 (2020), <https://arxiv.org/abs/2012.12453>
8. Jaspers, T.J.M., de Jong, R.L.P.D., Li, Y., Kusters, C.H.J., Bakker, F.H.A., van Jaarsveld, R.C., Kuiper, G.M., van Hillegersberg, R., Ruurda, J.P., Brinkman, W.M., Pluim, J.P.W., de With, P.H.N., Breeuwer, M., Al Khalil, Y., van der Sommen, F.: Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis* **108**, 103873 (2026). <https://doi.org/10.1016/j.media.2025.103873>

9. de Jong, R.L.P.D., Li, Y., Jaspers, T.J.M., van Jaarsveld, R.C., Kuiper, G.M., Badaloni, F., van Hillegersberg, R., Ruurda, J.P., van der Sommen, F., Pluim, J.P.W., Breeuwer, M.: Towards effective surgical representation learning with DINO models. In: MIDL (ed.) *Medical Imaging with Deep Learning* (2026), <https://openreview.net/forum?id=6FoIDPKzRV>
10. Jong, M.R., Boers, T.G., Fockens, K.N., et al.: Gastronet-5m: A multicenter dataset for developing foundation models in gastrointestinal endoscopy. *Gastroenterology* **170**(1), 174–187 (2026). <https://doi.org/10.1053/j.gastro.2025.07.030>
11. Kerssies, T., Cavagnero, N., Hermans, A., Norouzi, N., Averta, G., Leibe, B., Dubbelman, G., de Geus, D.: Your ViT is Secretly an Image Segmentation Model. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25303–25313 (2025). <https://doi.org/10.1109/CVPR52734.2025.02356>
12. Maier-Hein, L., Eisenmann, M., Sarikaya, D., et al.: Surgical data science – from concepts toward clinical translation. *Medical Image Analysis* **76**, 102306 (2022). <https://doi.org/10.1016/j.media.2021.102306>
13. Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., et al.: Heidelberg colorectal data set for surgical data science in the sensor operating room. *Scientific data* **8**(1), 101 (2021)
14. Mascagni, P., Alapatt, D., Murali, A., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., Padoy, N.: Endoscopes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Scientific Data* **12**(1), 331 (2025). <https://doi.org/10.1038/s41597-025-04642-4>
15. Mascagni, P., Vardazaryan, A., Alapatt, D., et al.: Artificial intelligence for surgical safety: Automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Annals of Surgery* **275**(5) (May 2022)
16. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: Vspw: A large-scale dataset for video scene parsing in the wild. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4133–4143 (June 2021)
17. Norouzi, N., Zulfikar, I., Cavagnero, N., Kerssies, T., Leibe, B., Dubbelman, G., de Geus, D.: VidEoMT: Your ViT is Secretly Also a Video Segmentation Model. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (to appear) (2026)
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024), <https://openreview.net/forum?id=a68SUt6zFt>, featured Certification
19. Schmidgall, S., Kim, J.W., Jopling, J., Krieger, A.: General surgery vision transformer: A video pre-trained foundation model for general surgery (2024), <https://arxiv.org/abs/2403.05949>
20. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3 (2025), <https://arxiv.org/abs/2508.10104>

21. Twinanda, A.P., Shehata, S., et al.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging* **36**(1), 86–97 (2017). <https://doi.org/10.1109/TMI.2016.2593957>
22. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 101–111. Springer (2023)
23. Xiong, X., Wu, Z., Lu, L., Xia, Y.: Sam3-unet: Simplified adaptation of segment anything model 3 (2025), <https://arxiv.org/abs/2512.01789>
24. Xiong, X., Wu, Z., Tan, S., Li, W., Tang, F., Chen, Y., Li, S., Ma, J., Li, G.: Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *Visual Intelligence* **4**(1), 2 (2026)
25. Zia, A., Bhattacharyya, K., Liu, X., et al.: Surgical tool classification and localization: results and methods from the miccai 2022 surgtoolloc challenge (2023), <https://arxiv.org/abs/2305.07152>