



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.

Stitch-Inferencer: Enhance Endoscopic Video Segmentation and Tracking via Panoramic Reconstruction

Shunsuke Kikuchi¹[0009–0008–3353–657X], Atsushi Kouno¹[0000–0002–6365–7638],
and Hiroki Matsuzaki¹

Jmees Inc, 5-4-6 Kashiwanoha, Kashiwa-city, Chiba, Japan
engineer@jmees-inc.com

Abstract. Surgical video understanding is fundamental to navigation systems. Endoscopic perception is often hindered by a limited field-of-view and frequent instrument occlusions, making spatio-temporal context essential for robust inference. These challenges have motivated video models that aggregate information across frames. However, existing video models typically store past observations implicitly in learned feature representations, often requiring task-specific video training, substantial annotated data, and increased computational cost. We propose **Stitch-Inferencer**, a real-time, model-agnostic inference framework that replaces implicit feature memory with an explicit image-space panoramic canvas. By stitching valid observations across frames, Stitch-Inferencer preserves previously observed pixels in an online, instrument-free view, expanding the effective field-of-view and providing direct access to regions that are temporarily occluded or absent from the current frame. Downstream segmentation or tracking models are applied to a compact region of interest on the panorama, and their predictions are reprojected to the current frame, enabling existing models to exploit long-range context without retraining. Experiments on anatomy segmentation and point/box tracking demonstrate consistent improvements across diverse baselines while preserving real-time throughput. The stitching module alone runs at over 60 FPS, providing a practical inference-time solution to enhance endoscopic perception in computationally constrained intra-operative environments. All source code and model weights are available at <https://github.com/JmeesInc/Stitch-seg.git>.

Keywords: Surgical video understanding · Image stitching · Segmentation

1 Introduction

Endoscopic perception in intraoperative settings is often characterized by partial observability and tight latency requirements, especially for surgical navigation and assistance systems that must update reliably in real time. Clinically relevant anatomy may lie outside the narrow field-of-view, be temporarily occluded by

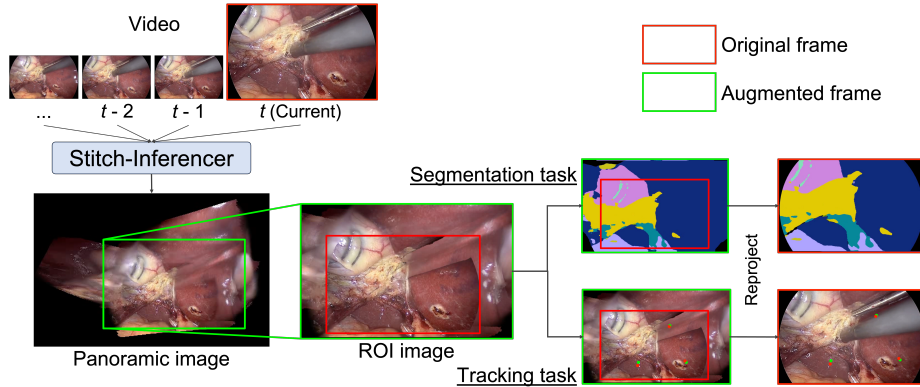


Fig. 1. Graphical abstract. Stitch-Inferencer creates an instrument-free panoramic view and improves downstream segmentation and tracking at inference time without additional training; green and red boxes indicate the panorama context and the current-frame region with reprojection, respectively.

instruments, or only become interpretable when compared against earlier observations. While per-frame recognition, segmentation, and tracking have advanced rapidly, relying on the current frame alone remains brittle in realistic procedures, motivating the need for long-range context in surgical video understanding.

A dominant approach to incorporating such context is to learn temporal memory through recurrent architectures, temporal attention, spatio-temporal convolution, or recent sequence models [27, 6, 22]. Although effective, these methods typically store past observations implicitly in latent feature representations. As a result, previously visible anatomical regions are not explicitly preserved once they leave the current field-of-view or become occluded; the model must instead rely on compressed feature states to recover or infer the missing context. This implicit form of memory can be powerful when trained with sufficient data, but it also increases architectural complexity and often requires task-specific training to exploit temporal information effectively.

We revisit a simple and comparatively underexplored alternative: maintaining memory directly in the image space. Recent advances in image matching make robust 2D alignment increasingly practical [34, 14]. In a large class of laparoscopic procedures, camera motion is moderate over short windows and the view is often approximately perpendicular to the anatomical surface. This makes it possible to accumulate previously observed anatomy into an online panoramic canvas. Unlike latent feature memory, such a canvas preserves observed pixels explicitly, expands the effective field-of-view, and provides direct access to regions that may be temporarily hidden by instruments or absent from the current frame. In this sense, the panorama acts as an explicit spatial memory for downstream inference.

Based on this idea, we propose **Stitch-Inferencer**, a real-time, model-agnostic framework that constructs an online, instrument-free panoramic view and per-

forms inference on the panorama with projection back to the current frame. At each time step, invalid regions such as image borders, instruments, and trocar ports are masked, valid observations are aligned and composited into a global canvas, and a compact region-of-interest around the current frame is extracted for downstream inference. Existing segmentation or tracking models can then be applied directly to this augmented view without retraining, and their predictions are reprojected to the original frame coordinates. This design upgrades single-frame models with temporally accumulated context at inference time while preserving online operation.

We evaluate Stitch-Inferencer on anatomy segmentation and point/box tracking and observe consistent improvements across diverse baselines while preserving real-time inference speed. Our contributions are: (1) an online method for constructing an interpretable explicit image-space spatial memory from endoscopic video, producing an instrument-free panoramic view robust to common surgical artifacts; (2) a model-agnostic projection/inverse-projection inference module that enhances existing models with long-range context at inference time, without retraining; and (3) empirical gains on segmentation and tracking under real-time constraints.

2 Methods

Stitch-Inferencer maintains an online instrument-free panoramic view and enables inference on the panorama with projection back to the current frame. The module is a lightweight state machine defined by a canvas image C_t , a canvas valid mask V_t , and a cumulative warp G_t . Given frame $I_t \in \mathbb{R}^{h \times w \times 3}$, we allocate a canvas of size $3h \times 3w$ (optionally scaled). Let T translate the frame origin to the canvas center. We denote image warping by $\mathcal{W}(I; G)$ (bilinear for RGB, nearest for masks). At each time step, the pipeline in Fig. 2 proceeds in five stages: (i) mask invalid regions; (ii) estimate the inter-frame homography from matched keypoints; (iii) update the cumulative warp and derive the frame–canvas mapping; (iv) warp and composite the frame into the canvas; (v) run a downstream model on the canvas and reproject it to the current frame.

Invalid-region mask. We form an invalid mask $M_t = M_t^{\text{border}} \cup M_t^{\text{tool}} \cup M_t^{\text{port}}$. M_t^{border} is obtained by 8-bit grayscale thresholding of extremely dark pixels (≤ 15), fitting an ellipse to the visible circular field-of-view, and marking pixels outside the ellipse as invalid. M_t^{tool} and M_t^{port} are produced by ConvNeXt-tiny U-Net models [24, 15] trained on multi-dataset surgical tool annotations [1, 10, 36, 19, 25, 23] and on Cholec80-port [13] (with label alignment on m2caiSeg [18] and GynSurg [20]) using the same image-training protocol as anatomy segmentation baselines (input 512×512 , AdamW [16], learning rate 5×10^{-5} , affine, blur and photometric augmentations). We define a masked frame $\tilde{I}_t = I_t \odot (1 - M_t)$.

Matching and inter-frame homography. We detect and describe keypoints with ALIKED [34] and match with LightGlue [14]. Keypoints falling inside M_t are discarded. For tracking experiments only, we additionally apply flow-based keypoint filtering (particularly important for tracking, which is more

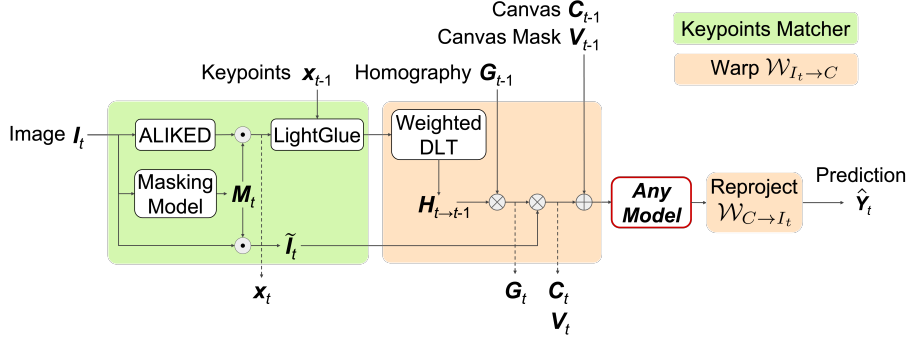


Fig. 2. Overview of the Stitch-Inferencer pipeline. Green: keypoint matching and homography estimation. Orange: warping, canvas update, and reprojection.

sensitive to pixel-level misalignment). We compute dense optical flow using NeuFlowv2 [33] and discard keypoints whose flow magnitude is a 2σ outlier after standardization: letting $m(\mathbf{x}_t)$ denote the flow-magnitude map, we compute its mean μ_t and standard deviation σ_t over valid pixels and discard keypoints at locations $\mathbf{x}_{t,i}$ satisfying $|m(\mathbf{x}_{t,i}) - \mu_t| > 2\sigma_t$.

From the remaining correspondences $\{(\mathbf{x}_{t,i}, \mathbf{x}_{t-1,i}, w_i)\}$ where match confidence weight w_i is given by LightGlue, we estimate $H_{t \rightarrow t-1}$ using weighted Direct Linear Transform (DLT). If fewer than four matches remain, we re-initialize (C_t, V_t, G_t) from I_t .

Cumulative warp and frame-to-canvas mapping. We update the cumulative warp $G_t = H_{t \rightarrow 0} = H_{t \rightarrow t-1} G_{t-1}$ with $G_0 = \mathbf{I}_3$. The current frame-to-canvas warp is $\mathcal{W}_{I_t \rightarrow C}(I) := W(I; S T G_t)$, and its inverse mapping is $\mathcal{W}_{C \rightarrow I_t} := \mathcal{W}_{I_t \rightarrow C}^{-1}$, where S is an optional scale matrix and T centers the current frame on the canvas. If the implied update is extreme (shear $> 15^\circ$, rotation $> 15^\circ$, or scale > 2.0), we re-anchor the panorama by re-centering the canvas on the current frame, set $G_t = \mathbf{I}_3$, and restart accumulation from this frame.

Compositing and Update States. We compute a Laplacian-variance blur score b_t on the current frame and skip the storage update when $b_t < \tau_{\text{blur}}$ ($\tau_{\text{blur}} = 30$), i.e., we keep $(C_t, V_t) = (C_{t-1}, V_{t-1})$ for the stored panorama. If $b_t \geq \tau_{\text{blur}}$, we update the canvas by warping the masked frame and alpha-blending only on valid pixels:

$$\tilde{I}_t^C = \mathcal{W}_{I_t \rightarrow C}(\tilde{I}_t), \quad V_t^{\text{new}} = \mathcal{W}_{I_t \rightarrow C}(1 - M_t), \quad V_t = V_{t-1} \cup V_t^{\text{new}} \quad (1)$$

$$C_t = \alpha_t \odot C_{t-1} + (1 - \alpha_t) \odot \tilde{I}_t^C \quad (\text{applied only where } V_t^{\text{new}} = 1). \quad (2)$$

where $\alpha_t^C = \mathcal{W}(\alpha_t^I; H_{t \rightarrow C})$ is a seam-aware blending weight. To smooth blending seams, we compute a spatial alpha-blending weight α_t based on a distance transform from the combined boundary of the invalid mask M_t and the frame edges. α_t decays linearly up to $r = 201$ pixels from the boundary (computed via iterative morphological dilation at $1/8$ scale).

Inference ROI and reprojection. For inference, we form an inference canvas by compositing the current masked frame onto the latest stored C_t , then crop a region of interest (ROI) that (i) contains the entire current frame content in canvas coordinates and (ii) lies within the valid stitched region V_t . The ROI is resized to the model input and we obtain predictions in ROI coordinates. For segmentation, we paste the ROI prediction back to the original ROI location on the canvas (zero elsewhere) to obtain $\hat{Y}_t^C$, then reproject to the current frame by $\hat{Y}_t = \mathcal{W}_{C \rightarrow I_t}(\hat{Y}_t^C)$. For tracking, we map points by $\hat{\mathbf{y}}_t \sim (S^T G_t)^{-1} \hat{\mathbf{y}}_t^C$, followed by homogeneous normalization.

3 Experiments and Results

We evaluated Stitch-Inferencer on two downstream tasks: anatomy segmentation and tracking. For both tasks, we compared original-frame inference with the same models augmented by Stitch-Inferencer, without any fine-tuning. For anatomy segmentation, we additionally compared against temporal and specialized video segmentation baselines. We also provide qualitative examples and ablate the inference-view construction to analyze how panoramic context should be provided to downstream models. Segmentation training and tracking experiments were conducted on an NVIDIA Quadro RTX 8000 with an AMD EPYC 7702P CPU, and segmentation inference speed was measured on an NVIDIA GeForce RTX 4090 with an Intel Core i9-10980XE CPU.

3.1 Anatomy segmentation

We evaluated anatomy segmentation on the CholecSeg8k [7] dataset. Since CholecSeg8k frame indices were not perfectly aligned with the original Cholec80 videos, we first matched frames to Cholec80 timestamps to enable evaluation under video inference. We used 5-fold cross-validation: rare-class clips (videos 01, 09, 12) were split with stratified K -fold at clip level, and all others used stratified group K -fold grouped by Cholec80 video to minimize leakage.

We tested Stitch-Inferencer on five single-frame baselines (Table 1) under a unified training protocol, and followed the original settings for additional temporal/specialized baselines. We reported macro Dice excluding background and

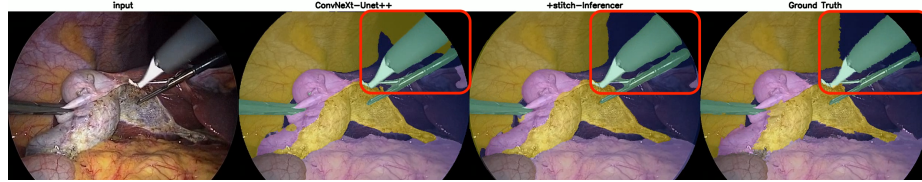


Fig. 3. Qualitative anatomy segmentation example on CholecSeg8k with ConvNeXt-B UNet++. From left to right: input frame, original inference, Stitch-Inferencer inference, and ground truth. Red boxes highlight a peripheral region where Stitch-Inferencer better recovers the ground-truth segmentation by leveraging panoramic spatial context.

Table 1. Video segmentation results with and without panorama inference. “Stitch” indicates inference on the stitched panorama with projection back to the current frame.

Method	Encoder	Stitch	Dice $\uparrow$	peri-Dice $\uparrow$	FPS $\uparrow$	p95 latency $\downarrow$
UNet++ [35]	ConvNeXt-B	✓	0.6884 0.7436	0.7159 0.7541	205.89 49.35	4.921 20.598
UPerNet [30]	ConvNeXt-B	✓	0.7100 0.7195	0.7489 0.7623	212.17 47.10	4.780 21.522
SegFormer [31]	MiT-B5	✓	0.7128 0.6995	0.7788 0.8008	104.89 45.03	10.136 22.512
DeepLabV3+ [5]	ResNet-101	✓	0.6813 0.7044	0.7566 0.7525	159.32 47.72	6.277 21.272
HRNet-seg [28]	HRNet-32	✓	0.6441 0.6508	0.7315 0.7362	119.74 51.72	8.816 19.651
SP-TCN [6]	HRNet-32		0.6016	0.6978	4.31	231.944
TMANet [27]	HRNet-32		0.5050	0.3257	9.47	105.556
DTERN [32]	MiT-B5		0.7006	0.6854	11.26	98.782
SALI [8]	PVTv2 [29]		0.6744	0.6976	14.65	70.394
UPerNet	VideoMamba [22]		0.7130	0.7517	13.44	74.853
MedicalSAM3 [9]	(box prompt)		0.5933	0.7256	2.55	393.933
SAM3 [3]	(point prompt)		0.3398	0.3257	2.28	442.724

tool classes (grasper, L-hook electrode), and additionally peri-Dice (outside a central ellipse, radius $> 2/3$). Speed was measured over 10,000 frames using FPS and p95 latency.

As shown in Table 1, panorama inference improved Dice for all evaluated single-frame baselines except SegFormer, without retraining or architectural changes. The mean Dice gain over sequences was $+0.0553 \pm 0.0174$ (UNet++), $+0.0096 \pm 0.0051$ (UPerNet), -0.0133 ± 0.0127 (SegFormer), $+0.0231 \pm 0.0071$ (DeepLabV3+), and $+0.0067 \pm 0.0046$ (HRNet-seg). Gains were also generally consistent on peri-Dice, indicating improved robustness under limited field-of-view. The stitched variants still ran at practical real-time throughput (45–52 FPS). Figure 3 provides qualitative examples of how the panoramic spatial memory changes downstream inference in dark region at screen edge. Overall, Stitch-Inferencer enhanced diverse segmentation models at inference time while maintaining practical throughput.

In addition to downstream inference speed, the stitching module alone (TensorRT) ran at 63.35 FPS with 16.049 ms p95 latency on the full Cholec80 dataset. Canvas re-initialization due to low matching occurred 4,207 times over 4,612,530 frames (0.092%), mainly in extra-corporeal scenes or when instruments occupied most of the view. The masks used for invalid-region removal achieved Dice scores of 0.86 for tool segmentation on EndoVis18 [2] and 0.77 for port segmentation on the Cholec80-port test set.

3.2 Tracking

We evaluated on STIR24 [26] (point tracking) and SurgT [4] (3D box tracking). For STIR24, we reported threshold accuracy average (δ -avg) at $\delta \in \{4, 8, 16, 32, 64\}$

Table 2. Comparison of tracking methods on STIR and SurgT datasets with and without Stitch-Inferencer.

Method	Stitch	STIR24		SurgT		
		δ -avg $\uparrow$	mean distance $\downarrow$	EAO $\uparrow$	Err2D $\downarrow$	Err3D $\downarrow$
CSRT [17]	✓	0.62246	44.229	0.552	6.0 ± 2.7	3.51 ± 3.8
		0.65908	25.807	0.579	5.7 ± 2.7	3.36 ± 2.5
Cotracker3 [11]	✓	0.70087	19.100	0.453	13.0 ± 11.0	3.13 ± 2.8
		0.71634	16.029	0.602	12.7 ± 11.1	2.79 ± 2.4
Litetracker [12]	✓	0.76298	17.150	0.585	5.1 ± 1.5	1.55 ± 1.1
		0.76215	14.906	0.685	5.2 ± 1.7	1.83 ± 1.3
MFT [21]	✓	0.77783	15.716	0.715	2.2 ± 1.3	1.47 ± 1.7
		0.77962	14.640	0.755	2.2 ± 1.2	1.39 ± 1.1

pixels and mean distance error (pixels). For SurgT, we tracked the 2D box center in left/right views, reconstructed 3D trajectories by stereo pairing, kept box size fixed to its initial value, and reported Expected Average Overlap (EAO), mean 2D error (px), and mean 3D error (mm). All trackers used public weights with no fine-tuning, and we kept the original video resolution.

Table 2 shows consistent gains across all evaluated trackers on STIR24 and SurgT without tracker-specific retraining. This supported Stitch-Inferencer as a model-agnostic inference module beyond segmentation: the temporally fused instrument-free panoramic view provided a shared spatial reference that improved robustness to tool occlusions and out-of-view motion. Qualitatively, targets often lost in the original frame view could be re-associated through the panorama, further indicating that these failure modes were addressed explicitly rather than by implicit temporal memorization.

3.3 Ablation study

A compact ablation on inference-view construction (Table 3). We compared the original per-frame baseline, the proposed *internal* ROI, an *external* ROI defined as the smallest bounding box that covered the valid stitched region ($V_t > 0$), and the *entire* canvas resized to the model input size. The proposed internal ROI performed best on both CholecSeg8k and STIR24, while the external and entire-canvas variants degraded performance. These results suggested that performance gains required not only panoramic context, but also an appropriately sized inference ROI: overly large views diluted relevant local detail and intro-

Table 3. Ablation of inference-view ROI. Results are reported on CholecSeg8k (Dice; UNet++) and STIR24 (δ -avg; MFT).

Metric	Baseline	Internal	External	Entire
CholecSeg8k Dice	0.6884	0.7436	0.6802	0.6718
STIR24 δ -avg	0.7778	0.7796	0.7789	0.4113

duced unnecessary background, whereas the proposed ROI preserved a compact, well-aligned context around the current frame.

4 Discussion and Conclusion

This work shows that an instrument-free panoramic view can serve as a practical, explicit spatial memory for endoscopic video understanding. Across anatomy segmentation and tracking, Stitch-Inferencer improved diverse baselines without retraining, supporting its role as a model-agnostic inference module. These results suggest that many common failure modes in endoscopic video understanding can be addressed by explicitly aggregating previously observed pixels in image space, rather than relying solely on implicit temporal memory in learned feature representations. Because the panorama is grounded in real observations rather than generative inpainting, it provides a reliable auxiliary view for surgical videos, where instrument occlusions, limited field-of-view, and intermittent visibility frequently degrade current-frame inference. Importantly, the method maintains real-time throughput while delivering measurable gains, making it compatible with intraoperative use cases such as navigation overlays and safety monitoring without requiring task-specific video-model training.

Interestingly, while CNN-based architectures consistently benefited from the expanded spatial context, SegFormer exhibited a slight performance drop. This may reflect architectural differences in context aggregation: CNN-based models rely more strongly on local feature composition and may exploit the expanded ROI, whereas SegFormer’s global self-attention may be more sensitive to appearance discontinuities introduced by stitching.

Limitations primarily stem from stitching assumptions and the current evaluation ecosystem. Planar homography remains vulnerable to strong non-planar deformation and parallax; when deformable tissue moves significantly—especially exiting the field-of-view—there is no mechanism to correct the accumulated canvas, yielding unnatural panoramas or stale context. In addition, recent public benchmarks provide limited coverage of long procedures with highly dynamic camera motion, constraining evaluation under practical regimes. Similar dataset limitations with densely labeled long videos make it difficult to assess if these gains persist when stronger models are trained with substantially more data.

Several future directions remain. Fine-tuning models on panorama ROIs may further improve accuracy. Since stitching can suppress arbitrary occluders given a mask, the instrument-free panoramic view may support annotation and higher-level endoscopic video understanding tasks (e.g., phase recognition and surgical VQA). Improving robustness under deformation (e.g., confidence-aware warping/blending, uncertainty estimation, or adaptive reset policies) is an important next step for reliable deployment.

Disclosure of Interests. Shunsuke Kikuchi and Atsushi Kouno are employees of Jmees Inc. Hiroki Matsuzaki is the co-founder and CEO of Jmees Inc.

References

1. Alabi, O., et al.: Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery. *Scientific Data* **12**(1), 825 (2025)
2. Allan, M., et al.: 2018 robotic scene segmentation challenge (2020), <https://arxiv.org/abs/2001.11190>
3. Carion, N., et al.: Sam 3: Segment anything with concepts (2025), <https://arxiv.org/abs/2511.16719>
4. Cartucho, J., et al.: Surgt challenge: Benchmark of soft-tissue trackers for robotic surgery. *Medical Image Analysis* **91**, 102985 (2024)
5. Chen, L.C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *ECCV*. pp. 833–851 (2018)
6. Grammatikopoulou, M., et al.: A spatio-temporal network for video semantic segmentation in surgical videos. *IJCARS* **19**(2), 375–382 (2024)
7. Hong, W.Y., et al.: Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80 (2020), <https://arxiv.org/abs/2012.12453>
8. Hu, Q., et al.: Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In: *MICCAI* (2024)
9. Jiang, C., et al.: Medical sam3: A foundation model for universal prompt-driven medical image segmentation (2026), <https://arxiv.org/abs/2601.10880>
10. Ju, X., et al.: Segcol challenge: Semantic segmentation for tools and fold edges in colonoscopy data (2024), <https://arxiv.org/abs/2412.16078>
11. Karaev, N., et al.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: *ICCV*. pp. 6013–6022 (2025)
12. Karaoglu, M.A., Ji, W., Abbas, A., Navab, N., et al.: Litetracker: Leveraging temporal causality for accurate low-latency tissue tracking. In: *MICCAI* (2025)
13. Kikuchi, S., et al.: Cholec80-port: A geometrically consistent trocar port segmentation dataset for robust surgical scene understanding (2026), <https://arxiv.org/abs/2602.17060>
14. Lindenberger, P., et al.: LightGlue: Local Feature Matching at Light Speed. In: *ICCV* (2023)
15. Liu, Z., et al.: A convnet for the 2020s. *CVPR* (2022)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *ICLR* (2019)
17. Lukežič, A., et al.: Discriminative correlation filter tracker with channel and spatial reliability. *International Journal of Computer Vision* (2018)
18. Maqbool, S., et al.: m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks. *CoRR* (2020)
19. Mascagni, P., et al.: Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Sci. Data* **12** (2025)
20. Nasirihaghighi, S., et al.: Gynsurg: A comprehensive gynecology laparoscopic surgery dataset (2025), <https://arxiv.org/abs/2506.11356>
21. Neoral, M., et al.: MFT: Long-term tracking of every pixel. In: *WACV*. pp. 6837–6847 (2024)
22. Park, J., et al.: Videomamba: Spatio-temporal selective state space model. In: *ECCV*. pp. 1–18 (2024)
23. Psychogyios, D., et al.: Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge (2024), <https://arxiv.org/abs/2401.00496>
24. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: *MICCAI*. pp. 234–241 (2015)

25. Rueckert, T., et al.: Video dataset for surgical phase, keypoint, and instrument recognition in laparoscopic surgery (phakir). arXiv:2511.06549 (2025), <https://arxiv.org/abs/2511.06549>
26. Schmidt, A., et al.: Surgical tattoos in infrared: A dataset for quantifying tissue tracking and mapping. *IEEE TMI* **43**, 2634–2645 (2024)
27. Wang, H., et al.: Temporal memory attention for video semantic segmentation. In: *ICIP*. pp. 2254–2258. *IEEE* (2021)
28. Wang, J., et al.: Deep high-resolution representation learning for visual recognition. *TPAMI* (2019)
29. Wang, W., et al.: Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media* **8**(3), 415–424 (2022)
30. Xiao, T., et al.: Unified perceptual parsing for scene understanding. In: *ECCV*. pp. 432–448 (2018)
31. Xie, E., et al.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: *NeurIPS* (2021)
32. Xu, X., et al.: Dual-temporal exemplar representation network for video semantic segmentation. In: *ICCV*. pp. 10775–10785 (2025)
33. Zhang, Z., et al.: Neuflow v2: Push high-efficiency optical flow to the limit (2025)
34. Zhao, X., et al.: Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation & Measurement* **72**, 1–16 (2023)
35. Zhou, Z., et al.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging* (2019)
36. Zia, A., et al.: Intuitive surgical surgtoolloc challenge results: 2022-2023 (2025), <https://arxiv.org/abs/2305.07152>