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.

STAG: A Spatio-Temporal Affinity Framework for Surgical Video Understanding

Alireza Kheyrkhah[1], Amir Hossein Saleknia[1], Sanaz Karimijafarbigloo[2], Reza Azad[2], Dorit Merhof[2], and Alaa Sulaiman[3]

[1]Independent Researcher, [2]University of Regensburg, Germany, [3]Abdora INC, USA
alaa@abdora.ai

Abstract. A frozen surgical foundation model such as LemonFM provides a strong backbone for endoscopic perception, yet the standard frozen backbone probing protocol discards both the spatial layout of the backbone's feature map and the temporal structure already encoded in its learned embedding geometry. Self-supervised distillation trains these embeddings so that cosine similarity reflects semantic similarity between frames, suggesting that a useful notion of temporal continuity already exists in the pretrained feature manifold and can be exploited directly rather than relearned from scarce downstream annotations. We introduce **STAG**, a compact *Spatio-Temporal Affinity Graph* that builds on this insight with three lightweight modules operating on the frozen backbone's learned geometry: it recovers spatial evidence through class-conditional slot pooling over the pre-pool token grid, replaces the classification head with a cosine prototype layer aligned with the backbone's metric space, and performs non-parametric temporal label propagation over a video graph whose edges are the backbone's own cosine similarities. Across four benchmark settings (Cholec80 phase recognition, Cholec80 and GraSP tool detection, and CholecT50 verb recognition), STAG surpasses the LemonFM baseline on every task, improving accuracy by 9.5 pp on phase recognition and mAP by 14.4, 5.2, and 6.0 pp on Cholec80 tool, GraSP tool, and CholecT50 verb detection respectively, while adding only 1-2 M parameters and keeping the backbone fully frozen. An ablation attributes these gains to each component: the spatial and prototype components drive large gains across all tasks, while temporal propagation adds +6.5 pp on densely sampled Cholec80 phase recognition. On GraSP, whose keyframes are 35 s apart with minimal continuity, the temporal component contributes -0.2 pp (gains arise from spatial and prototype modules), confirming that temporal reasoning engages where continuity is present. Compared to MS-TCN, STAG outperforms learned parametric convolution on tool detection by wide margins.

Keywords: Surgical Video, Spatio-Temporal Affinity, Label Propagation, Phase Recognition, Tool Detection, Action Recognition

1 Introduction

Surgical perception including phase recognition, instrument detection, and action understanding is critical for assistive and autonomous interventions. Because annotating surgical video requires clinical expertise, recent work uses self-supervised pretraining followed by lightweight adaptation [1]. The standard test of feature quality is to freeze the backbone and fit a per-frame classifier on the pooled feature [2], which assumes a globally pooled vector exposes everything a downstream task needs. This leaves two properties unexploited: most encoders retain a pre-pool spatial grid of tokens carrying localized tool and interaction evidence that pooling discards, and self-supervised distillation objectives like DINO [3] explicitly train cosine similarity between embeddings to reflect semantic similarity. In LemonFM [1], which distils a ConvNeXt-L encoder on ∼4k curated endoscopic videos, the frozen feature manifold already behaves as a similarity kernel across adjacent frames, yet a per-frame classifier has no mechanism to query that kernel.

Existing remedies address these gaps separately, discarding the backbone's geometry. Temporal modelling in surgical video relies on parametric predictors trained from scratch: MS-TCN [4] and TeCNO [5] learn causal dilated convolutions, Trans-SVNet [6] uses temporal attention, and SurgFUTR [7] and SurgTEMP [8] inject further priors, each rediscovering temporal structure that a distilled backbone may already encode. Slot attention [9] and DETR-style set prediction [10] extract class-specific evidence from spatial token grids, but only end-to-end on general-purpose detection benchmarks.

LemonFM's DINO-style geometry suggests a more direct alternative: a downstream head can query the continuity already present in the pretrained embedding manifold. We propose **STAG**, a *Spatio-Temporal Affinity Graph* stacking three lightweight components on a frozen backbone: (i) a class-conditional slot pool cross-attending learnable queries to the spatial token grid, restoring localized evidence; (ii) a cosine prototype layer whose logits inhabit the backbone's distilled metric geometry; and (iii) a temporal affinity graph propagating per-frame logits over a video graph whose edges are the backbone's own cosine similarity, with only two scalars learned.

We evaluate STAG on Cholec80 [11] phase and tool recognition, GraSP [12] instrument detection, and CholecT50 [13] verb recognition, using LemonFM as the frozen backbone. STAG exceeds the LemonFM baseline on every task, adding only 1-2 M parameters. An ablation study over the three components shows that the spatial and prototype components drive large gains everywhere, while the temporal component adds improvement where temporal continuity exists. Compared to MS-TCN on the same frozen features, reusing the backbone's similarity as a temporal kernel outperforms relearning it from scratch on spatially demanding tool-detection tasks.

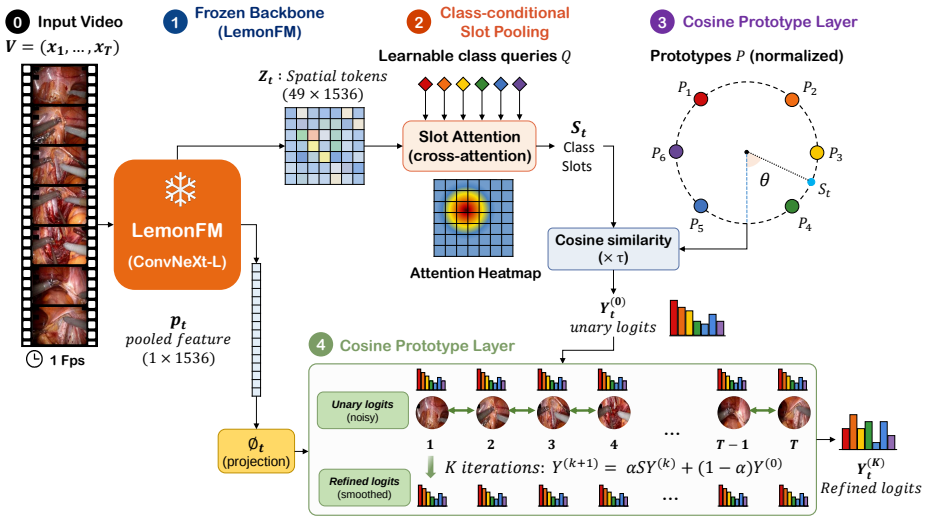


Fig. 1: Overview of STAG. The frozen LemonFM backbone emits a pooled feature p_t and a 7×7 token grid Z_t per frame; slot pooling (orange) and cosine prototypes yield unary logits $Y_t^{(0)}$ (blue), which label propagation on the p_t similarity graph refines into temporally consistent predictions $Y_t^{(K)}$ (green).

2 Method

Fig. 1 illustrates the full STAG pipeline. A frozen LemonFM backbone produces two outputs per frame: a global pooled feature p_t and a spatial token grid Z_t, which feed three lightweight modules that progressively refine frame-level predictions into temporally consistent logits.

2.1 Overview

Let $\mathcal{V} = (x_1, \ldots, x_T)$ denote a surgical video of T frames sampled at 1 fps, and let g_θ be the frozen LemonFM teacher backbone [1]. We hook the pre-average-pool activation to obtain per-frame outputs

$$(p_t,\ Z_t) = g_\theta(x_t), \qquad p_t \in \mathbb{R}^{1536},\ Z_t \in \mathbb{R}^{49\times 1536}, \tag{1}$$

where p_t is the pooled feature used in LemonFM's baseline evaluation and Z_t is the 7×7 spatial token grid. STAG feeds Z_t to a spatial stage and p_t to a temporal one, returning per-frame logits $Y_t \in \mathbb{R}^{N_c}$ for N_c classes. Backbone parameters are never updated; the trainable head is at most 2M parameters.

Class conditional slot pooling. STAG first recovers spatial evidence from the token grid. We introduce K_s learnable class queries $Q \in \mathbb{R}^{K_s \times d}$ $(d = 256)$ and

cross-attend them once to the 49 spatial tokens:

$$S_t = \mathrm{LN}\Big(\mathrm{MHA}\big(Q,\ W_k Z_t,\ W_v Z_t\big)\Big)\ \in\ \mathbb{R}^{K_s \times d}. \tag{2}$$

For phase recognition $K_s = 1$: a single context slot pools the whole scene. For tool and verb recognition $K_s = N_c$, so query k is a class-specific detector. Unlike iterative Slot Attention [9], a single cross-attention pass suffices since the queries attend to a frozen, already-semantic grid; the attention weights act as soft class-specific masks over the 7×7 grid (Fig. 1), recovering spatial evidence that global pooling discards.

Cosine prototype layer. These slots are turned into logits in the backbone's own metric space. LemonFM's DINO projection head is itself a normalized cosine classifier over 65,536 prototypes [3]; to place our task classes inside that same geometry, we learn prototypes $P \in \mathbb{R}^{N_c \times d}$ and produce logits via temperature-scaled cosine similarity:

$$Y^{(0)}_{t,c} = \tau \cdot \frac{S^{(c)}_t}{\|S^{(c)}_t\|} \cdot \frac{P_c}{\|P_c\|}, \qquad \tau = 20. \tag{3}$$

When $K_s = 1$, the single slot is compared against all N_c prototypes; when $K_s = N_c$, slot c is compared only with prototype c. The temperature $\tau = 20$ rescales cosine similarities to the range cross-entropy and BCE expect; prototypes are initialized near zero and learned jointly with the slot queries, replacing the implicit Euclidean assumption of a standard classification layer with metric learning on the same sphere the features were distilled on.

Temporal affinity graph. Finally, per-frame logits are smoothed over time by STAG's most distinctive component, a non-parametric temporal aggregator exploiting LemonFM's own similarity kernel. Let $\varphi_t = W_\varphi p_t / \|W_\varphi p_t\|$ be a normalized projection, and define a symmetric affinity restricted to a temporal window W:

$$A_{ij} = \exp\Big(-\frac{\|\varphi_i - \varphi_j\|^2}{\sigma^2}\Big)\,\mathbf{1}\big[|i-j| \le W\big], \qquad S = D^{-1/2} A\, D^{-1/2}, \tag{4}$$

with $D = \mathrm{diag}(\sum_j A_{ij})$. We then iterate Zhou-style label propagation [14]:

$$Y^{(k+1)} = \alpha\, S\, Y^{(k)} + (1-\alpha)\, Y^{(0)}, \qquad k = 0, \ldots, K-1, \tag{5}$$

where $\alpha \in (0,1)$ balances propagated evidence against the original unary prediction. Because A is banded to $|i-j| \le W$, each iteration costs $O(TWN_c)$ instead of $O(T^2 N_c)$, smoothing an entire 1 fps video in one pass; symmetric normalization keeps the recursion stable. Only α and σ are learnable; the filter shape is inherited from LemonFM's pretraining rather than learned from scarce labels—contrast a TeCNO head [5], whose kernels ($\sim 10^5$–10^6 parameters) must rediscover surgical similarity structure from limited downstream data. Task specialization reduces to choosing W and K: for Cholec80 phase ($K_s = 1$, $W = 256$,

$K=8$) propagation covers long monotonic segments; for Cholec80 tool ($K_s=N_c$, $W=64$, $K=4$) it captures persistent presence; for GraSP tool ($K_s=N_c$, $W=8$, $K=2$) only a short window is meaningful since keyframes are 35 s apart; and for CholecT50 verb ($K_s=N_c$, $W=8$, $K=1$) propagation operates over transient bursts.

2.2 Training

STAG is trained per-video, with one full video per optimization step so the temporal graph spans the entire sequence, using AdamW (initial LR 3×10^{-4}, cosine schedule, gradient clipping at 1.0) with mixed precision on a single H100 GPU. Phase recognition uses softmax cross-entropy; tool and verb recognition use binary cross-entropy with logits for their multi-label structure and class imbalance. The backbone remains frozen throughout.

3 Experiments

3.1 Datasets, Protocols, and Implementation

We follow the LEMON paper's splits and metrics [1] across four settings spanning different temporal regimes. **Cholec80** [11]: 80 cholecystectomy videos (∼184k frames at 1 fps, 7 phases, 7 tools, 40/8/32 split), with long monotonic phases and interrupted tool presence. **GraSP** [12]: radical prostatectomy videos, 7 instruments, keyframes 35 s apart (4/4/5 split, ∼3.4k keyframes), minimal temporal continuity. **CholecT50** [13]: 10 transient verbs on every 1-fps frame of 50 videos, 5-fold cross-validated mAP. Phase uses accuracy/macro F1; tool and verb use macro mAP, robust to class imbalance.

The frozen backbone is LemonFM's teacher ConvNeXt-L [15]; inputs are resized to 224^2 with ImageNet normalization, and a forward hook on the pre-average-pool activation yields the pooled feature $p_t \in \mathbb{R}^{1536}$ and spatial map $Z_t \in \mathbb{R}^{49\times1536}$, precomputed once per dataset in fp16 with no test-time augmentation. Training uses AdamW (LR 3×10^{-4}, weight decay 10^{-4}, cosine schedule, 50 epochs, fixed seed), selecting the best validation checkpoint, with W, K set per task as in Sec. 2.

3.2 Main Results

Table 1 compares STAG against seven foundation models under LemonFM's standard per-frame classifier-head protocol, as reported in [1]; under the same backbone and splits, STAG surpasses these methods by wide margins on all four settings.

STAG improves on LemonFM by +9.5 pp accuracy / +9.6 pp F1 (Cholec80 phase), +14.4 pp mAP (Cholec80 tool), +5.2 pp mAP (GraSP tool), and +6.0 pp mAP (CholecT50 verb), the largest gain arising where localized spatial evidence matters most. STAG also exceeds every other backbone, including SurgeNetXL,

Table 1: Comparison with surgical foundation models under a frozen backbone. Results for all methods other than STAG are reproduced from Che et al. [1] using per-frame classifiers on shuffled minibatches; STAG is evaluated under the same backbone and splits.

Method	Pretrain	Phase		Tool		Action
		Cholec80		Cholec80	GraSP	CholecT50
		Acc	F1	mAP	mAP	mAP
MAE [16]	LemonFM	54.9	43.4	30.6	45.2	24.8
VideoMAEv2 [17]	LemonFM	55.8	48.5	39.6	55.9	25.8
Endo-FM [18]	Private	62.7	53.9	50.0	64.3	26.2
EndoViT [19]	Merged	46.9	39.0	34.7	47.1	29.5
GSViT [20]	GenSurg.	20.2	13.5	18.2	35.6	16.5
SurgeNetXL [21]	SurgNetXL	73.2	65.1	72.1	62.7	45.3
LemonFM [1]	LemonFM	<u>75.8</u>	<u>68.6</u>	<u>76.8</u>	<u>76.4</u>	<u>50.4</u>
STAG (ours)	LemonFM	**85.3**	**78.2**	**91.2**	**81.6**	**56.4**
Δ vs. LemonFM		**+9.5**	**+9.6**	**+14.4**	**+5.2**	**+6.0**

Phase: 7 classes (Cholec80). Tool: 7-class presence (Cholec80, GraSP). Action: 10 verbs (CholecT50, 5-fold, ±2.0). Best in **bold**, second best <u>underlined</u>; Δ = absolute gain.

by 5–15 pp despite only 1–2 M added parameters. Since the comparison holds the backbone fixed across all methods, STAG's gains isolate the value of exploiting the spatial token grid and the pretrained similarity kernel, rather than conflating improvements from backbone scale or pretraining data with architectural choices.

4 Analysis

We ablate STAG's modules on the *same* frozen LemonFM backbone under identical training (one video/step, AdamW, 50 epochs, seed 42), adding mechanisms sequentially: (a) plain classifier on pooled feature, (b) cosine prototype layer, (c) slot pooling on spatial tokens, and (d) temporal affinity graph. Table 2 and Fig. 2 report metrics and incremental gains on the tool detection tasks. Rows (b)–(c) add learned capacity, conflating geometry exploitation with richer heads; row (d) adds only two scalars (α, σ), directly testing temporal geometry reuse. Unlike Table 1, where LemonFM trains on shuffled minibatches, our ablation trains on full videos (no shuffling) to enable the temporal graph in row (d). This removes shuffling regularization, which lowers the sparse/transition-heavy baselines (GraSP $76.4 \rightarrow 63.4$, Cholec80 tool $76.8 \rightarrow 66.8$). Head† therefore serves as a matched anchor, isolating one mechanism per row.

Table 2: Cumulative module contributions of STAG on tool detection (each row adds one mechanism). Head† is the internal regime-matched anchor (Section 4), not the LemonFM baseline of Table 1. Best in **bold**; CholecT50 5-fold mean±std.

Head configuration	Tool mAP		Verb mAP
	Cholec80	GraSP	CholecT50
(a) Head† *(frame-wise baseline)*	66.8	63.4	40.1,±2.2
(b) + Prototype *(metric)*	80.9	79.1	50.3,±1.1
(c) + Slot pool *(spatial)*	88.9	81.8	54.0,±2.1
(d) + Temporal graph *(temporal)* = **STAG**	**91.2**	**81.6**	**56.4**,±2.0
total Δ over Head†	+24.4	+18.2	+16.3

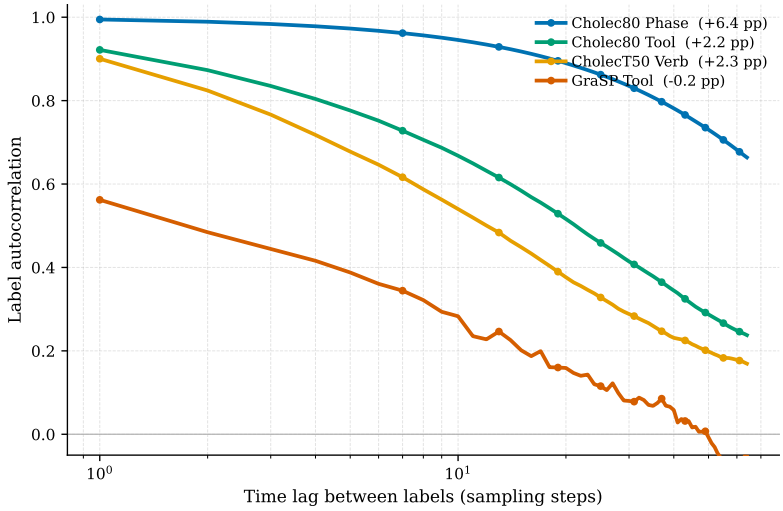


Fig. 2: Label autocorrelation versus temporal lag, measured from ground-truth test sequences; legend reports each task's temporal-module gain. Lag is in label sampling steps (1 s for 1 fps tasks, 35 s for GraSP keyframes).

Cosine prototype layer. The cosine prototype layer yields large gains on tool detection: +14.2 pp on Cholec80 and +15.7 pp on GraSP (and +10.2 pp on CholecT50 verb).

Slot pooling. Class-conditional queries on the 49-token spatial grid further improve tool detection by +8.0 pp on Cholec80 and add +3.6 pp on CholecT50 verb. Fig. 3 confirms that the queries localize instruments and interaction sites.

Temporal affinity graph. With only two learnable scalars, this module's gains reflect the frozen embedding's continuity. It adds +2.2 pp on Cholec80 tool and

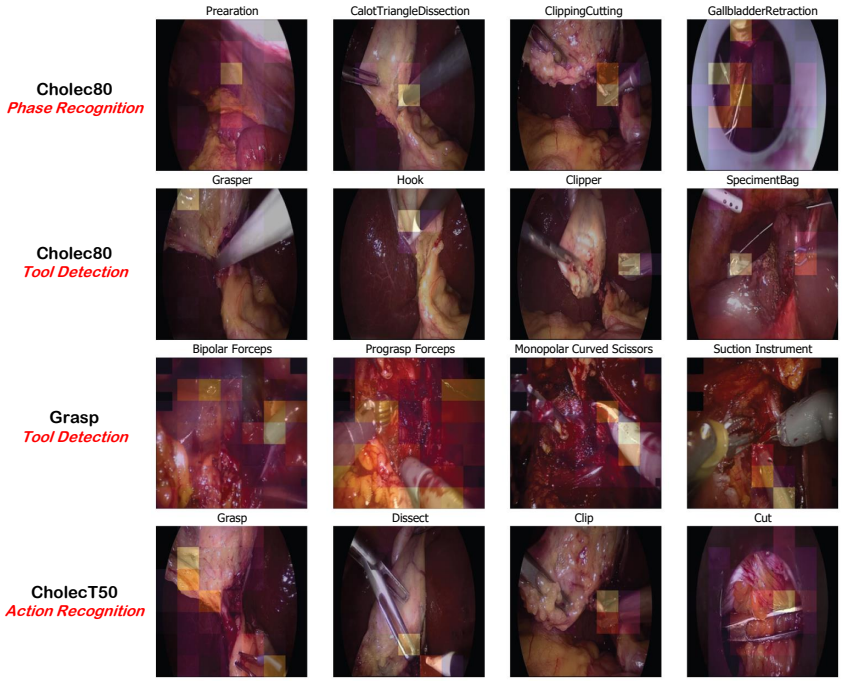


Fig. 3: Slot attention of trained STAG on the different tasks.

+2.4 pp on CholecT50 verb, but shows −0.2 pp on GraSP (minimal continuity due to 35 s keyframe spacing). Full STAG reaches **91.2** and **81.6 mAP** on Cholec80/GraSP tool detection.

Comparison with parametric temporal heads. To isolate STAG's non-parametric temporal module, we compare it against MS-TCN [4,5] trained on identical frozen LemonFM features under the same protocol. On tool detection (3), STAG substantially outperforms: +26.2 pp mAP on Cholec80 and +27.0 pp on GraSP. This confirms that reusing LemonFM's similarity kernel as a temporal prior captures more relevant structure than a learned TCN from the same frozen features.

Table 3: STAG vs. MS-TCN on frozen LemonFM features for tool detection. Both heads use the same cached features and training protocol.

Head	Params	**Tool (mAP)**	
		Cholec80	GraSP
MS-TCN [4]	0.12–0.76 M	65.1	54.6
STAG (ours)	1–2 M	**91.2**	**81.6**

5 Conclusion

We introduced STAG, a Spatio-Temporal Affinity Graph exploiting spatial and temporal structure already encoded in a frozen surgical foundation model: STAG recovers it through class-conditional slot pooling, a cosine prototype layer aligned with the backbone's metric space, and non-parametric temporal label propagation over the backbone's semantic kernel. Across four surgical benchmarks, STAG surpasses the LemonFM baseline by wide margins with only 1–2 M parameters, with temporal gains scaling with each task's measured continuity. Compared to MS-TCN on the same frozen features, reusing the pretrained similarity kernel outperforms relearning temporal structure on spatially demanding tool-detection tasks.

References

1. Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. 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*, pages 42659–42669, 2026.
2. Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, et al. Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis*, 88:102844, 2023.
3. Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
4. Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019.
5. Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *International conference on medical image computing and computer-assisted intervention*, pages 343–352. Springer, 2020.
6. Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In *International conference on medical image computing and computer-assisted intervention*, pages 593–603. Springer, 2021.
7. Saurav Sharma, Chinedu Innocent Nwoye, Didier Mutter, and Nicolas Padoy. State-change learning for prediction of future events in endoscopic videos. *arXiv preprint arXiv:2510.12904*, 2025.
8. Shi Li, Vinkle Srivastav, Nicolas Chanel, Saurav Sharma, Nabani Banik, Lorenzo Arboit, Kun Yuan, Pietro Mascagni, and Nicolas Padoy. Surgtemp: Temporal-aware surgical video question answering with text-guided visual memory for laparoscopic cholecystectomy. *arXiv preprint arXiv:2603.29962*, 2026.
9. Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf.

Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.
10. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
11. Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.
12. Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolás Fernández, et al. Pixel-wise recognition for holistic surgical scene understanding. *Medical Image Analysis*, page 103726, 2025.
13. Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.
14. Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.
15. Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.
16. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
17. Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023.
18. Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International conference on medical image computing and computer-assisted intervention*, pages 101–111. Springer, 2023.
19. Dominik Batić, Felix Holm, Ege Özsoy, Tobias Czempiel, and Nassir Navab. Endovit: pretraining vision transformers on a large collection of endoscopic images. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1085–1091, 2024.
20. Samuel Schmidgall, Ji Woong Kim, Jeffrey Jopling, and Axel Krieger. General surgery vision transformer: A video pre-trained foundation model for general surgery. *arXiv preprint arXiv:2403.05949*, 2024.
21. Ronald de Jong, HJ Carolus, HA Franciscus, Romy C van Jaarsveld, Richard van Hillegersberg, PW Josien, Peter HN de With, Yasmina al Khalil, Fons van Der Sommen, et al. Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis*, page 103873, 2025.