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.

Efficient Surgical Phase Recognition via Compact Transformer Encoding and Recurrent Temporal Modeling

Luning Li[1], Yohei Sanmoto[2], Yudai Goto[3], Xuhan Zhu[4], Qun Jin[5], Bo Peng[6], Zhiyong Zhou[6], YaKang Dai[6], Rong Xu[6⋆], and Xin Zhu[1⋆]

[1] Department of AI Technology Development, M&D Data Science Center, Institute of Biomedical Engineering, Institute of Science Tokyo, Tokyo, Japan
zhu.xin@tmd.ac.jp
[2] University of Tsukuba, Tsukuba, Japan
[3] University of Tsukuba Hospital, Tsukuba, Japan
[4] Graduate School of Medical and Dental Sciences, Institute of Science Tokyo, Tokyo, Japan
[5] Waseda University, Tokorozawa, Japan
[6] Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China
xurong@sibet.ac.cn

Abstract. Transformer-based methods have improved surgical phase recognition, but their large model size and high computational cost make them less practical for long surgical videos and resource-constrained settings. This paper investigates whether a lightweight student model can preserve temporal modeling ability while substantially reducing the parameter count and per-clip computation of SurgFormer. We propose a SF-Small+BiGRU framework that combines compact Transformer-based clip encoding with recurrent cross-clip temporal modeling. The resulting model reduces the parameter count from 177.8M to 31.9M and the per-clip computation from 893.28 to 150.24 GFLOPs. On Cholec80, the cross-entropy (CE)-trained model achieved 0.8956 accuracy and 0.7143 Jaccard, with a Jaccard decrease of 0.0175 compared with the SurgFormer teacher despite the substantial reduction in model size. On AutoLaparo, GRU-level distillation further improved Jaccard from 0.4997 to 0.5257. These results suggest that, under the evaluated settings, recurrent cross-clip temporal modeling contributes more consistently to the lightweight model than knowledge distillation, whose effects are dataset- and level-dependent.

Keywords: Surgical phase recognition · Knowledge distillation · Temporal modeling · BiGRU · Lightweight surgical phase recognition

⋆ Corresponding author.

1 Introduction

Surgical phase recognition aims to assign phase labels to surgical video frames or clips from intraoperative videos [3, 6, 11]. It turns long surgical videos into structured workflow information, supporting workflow analysis [2], automatic report generation [15], and future context-aware assistance [10].

Early methods relied on CNN-based frame-level representation learning and recurrent temporal modeling, such as EndoNet and SV-RCNet [14, 8], and were followed by stronger temporal modules such as TMRNet and TeCNO [9, 2]. More recent methods adopted Transformer-based spatio-temporal architectures, including Trans-SVNet, ARST, SurgFormer, and MoSFormer [5, 19, 17, 4], which improve recognition by modeling temporal dependencies within and across video clips. However, they are often large and compute-intensive: SurgFormer has about 178M parameters, and MoSFormer adds further overhead through multi-scale memory for long-range temporal reasoning. For long surgical videos, such cost increases training and offline inference time and may limit the practicality of large models in resource-constrained settings.

In this work, we propose a lightweight recognition framework that compresses SurgFormer into a SurgFormer-Small (SF-Small) encoder coupled with a bidirectional gated recurrent unit (BiGRU) [1] temporal module. The SF-Small encoder follows the Hierarchical Temporal Attention (HTA) architecture of SurgFormer, while the BiGRU explicitly aggregates cross-clip temporal context. Knowledge distillation [7] is used as an auxiliary training strategy rather than as the only source of model compression.

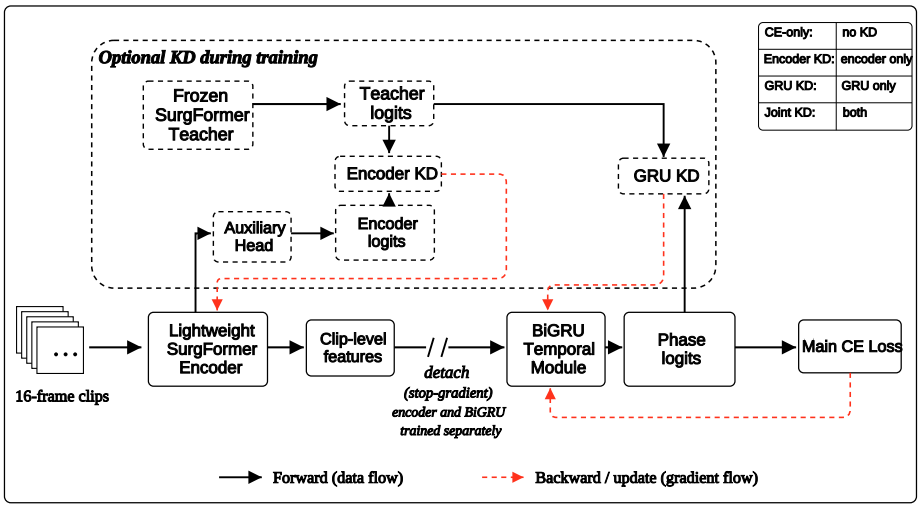
The main contributions of this work are as follows:

1. We design a SF-Small encoder with 384-dimensional clip embeddings, 8 Transformer layers, and 6 attention heads; with the BiGRU module, the final model has 31.9M parameters and 150.24 GFLOPs per clip.
2. We introduce a BiGRU temporal module that aggregates clip-level context bidirectionally across the whole video.
3. We compare encoder-level, GRU-level, and joint distillation, and find that BiGRU temporal modeling provides more consistent gains under the evaluated settings, while KD effects are dataset-dependent.

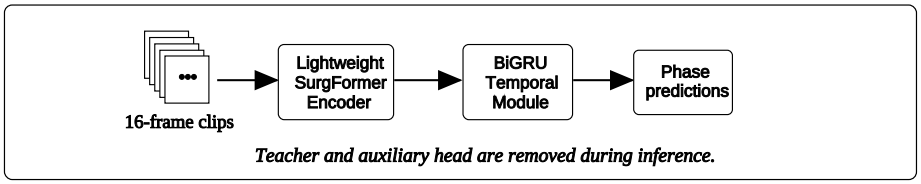
2 Method

2.1 Overall Framework

Figure 1 illustrates the proposed SF-Small+BiGRU framework. A surgical video is divided into non-overlapping 16-frame clips. Each clip is encoded by SF-Small, and the resulting clip-level features are passed to a BiGRU module for cross-clip temporal modeling and phase prediction. During training, an optional KD branch provides auxiliary supervision from a frozen SurgFormer teacher to the encoder output, the BiGRU output, or both; the teacher and auxiliary head are removed during inference.



(a) Training Stage



(b) Inference Stage

Fig. 1: Overview of the proposed SF-Small+BiGRU framework. (a) During training, an optional KD branch supervises the encoder or BiGRU output using a frozen SurgFormer teacher. (b) During inference, the teacher and auxiliary head are removed; only the lightweight encoder and BiGRU temporal module are retained.

2.2 SF-Small Encoder

SF-Small follows the HTA architecture of SurgFormer but reduces the embedding dimension from 768 to 384, the number of Transformer layers from 12 to 8, and the number of attention heads from 12 to 6. This results in a compact 29.9M-parameter encoder, compared with 177.8M for SurgFormer, that preserves intra-clip temporal modeling while reducing model complexity.

2.3 Temporal Modeling with BiGRU

Although HTA captures temporal dependencies within individual clips, it does not explicitly model temporal relationships across clips. We therefore introduce a BiGRU module after SF-Small to capture longer-range temporal context with modest additional cost.

Given a sequence of clip features extracted by SF-Small, a two-layer BiGRU with a hidden dimension of 256 and a dropout rate of 0.1 produces phase logits for each clip. The clip features are detached before being passed to the BiGRU, which decouples clip-level encoder training from video-level temporal modeling and reduces memory consumption. Because the BiGRU uses context from both preceding and subsequent clips, the framework is evaluated in an offline full-sequence setting rather than a strictly online one.

2.4 Training Objectives and Distillation Strategies

The main recognition branch predicts phases from the BiGRU output. Motivated by prior self-distillation studies in surgical phase recognition [18], we instead adopt a teacher–student setting, using a SurgFormer model as a frozen teacher and treating KD as an auxiliary strategy for training the lightweight student. Both branches are trained with the same combined objective applied at different locations. For logits z, hard labels y, and frozen teacher logits z^t,

$$\mathcal{L}(z;\alpha) = (1-\alpha)\,\mathrm{CE}(z,y) + \alpha T^2\,\mathrm{KL}\big(\sigma(z^t/T)\,\|\,\sigma(z/T)\big)\,, \tag{1}$$

where σ denotes the softmax and $\alpha = 0$ reduces $\mathcal{L}$ to plain cross-entropy. The encoder and its auxiliary head are trained with $\mathcal{L}(z^{\mathrm{enc}};\alpha_{\mathrm{enc}})$ on the per-clip auxiliary logits, and the BiGRU with $\mathcal{L}(z^{\mathrm{gru}};\alpha_{\mathrm{gru}})$ on the sequence logits; the four strategies differ only in which weight is non-zero. CE-only uses $\alpha_{\mathrm{enc}} = \alpha_{\mathrm{gru}} = 0$; encoder-level KD sets $\alpha_{\mathrm{enc}} = \alpha$; GRU-level KD sets $\alpha_{\mathrm{gru}} = \alpha$; and joint KD sets both to α. Because the clip features are detached, $\mathcal{L}(z^{\mathrm{gru}})$ never reaches the encoder, whose gradients come solely from $\mathcal{L}(z^{\mathrm{enc}})$. That term keeps a hard-label cross-entropy component in every strategy, so the encoder is supervised in all four settings. The auxiliary head is used only during training and is discarded at inference.

3 Experiments

We examined three questions: whether the lightweight model retained much of the SurgFormer teacher's performance, whether BiGRU improved cross-clip temporal modeling, and how different KD locations affected performance.

3.1 Datasets and Evaluation Metrics

Cholec80 [14] contains 80 laparoscopic cholecystectomy videos with seven annotated phases; following the TeCNO split [2], we used 40 videos for training, 8 for validation, and 32 for testing. AutoLaparo [16] contains 21 laparoscopic surgery videos with seven annotated phases; we followed the official split, using 10 videos for training, 4 for validation, and 7 for testing.

We reported frame-level accuracy, macro-F1, and macro-Jaccard, hereafter referred to as F1 and Jaccard for brevity. Since the model predicted one phase

label for each 16-frame clip, each clip-level prediction was expanded to its corresponding frames before evaluation. Metrics were computed over all pooled test frames without temporal relaxation or boundary tolerance. Because many prior works aggregate per video or per phase, these values are not directly comparable with published numbers and should be read as a controlled internal comparison.

3.2 Implementation Details

All experiments were conducted on a single NVIDIA RTX 5090 GPU with 32 GB of memory. Input frames were sampled at 1 FPS and resized to 224×224. Each video was divided into non-overlapping 16-frame clips with a stride of 16.

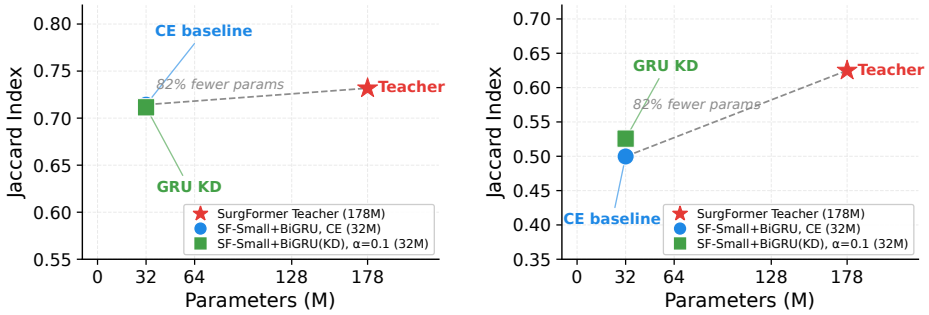
The SurgFormer teacher was trained by us separately on each dataset under the same splits and clip sampling as the student, rather than taken from a released checkpoint; its logits were pre-computed offline on exactly the same non-overlapping 16-frame clips and cached, so teacher and student targets are clip-aligned by construction. The SF-Small encoder was initialized from ImageNet-pretrained ViT-Small weights (`vit_small_patch16_224`). The encoder and its auxiliary classification head were optimized with AdamW at a learning rate of 1×10^{-5}, and the BiGRU module separately with AdamW at 1×10^{-4}; both used a weight decay of 1×10^{-4} and cosine annealing to a minimum learning rate of 1×10^{-6}. Training ran for at most 50 epochs, with early stopping on the validation Jaccard index and a patience of 10 epochs.

For knowledge distillation, we set the temperature to $T = 4$ and the balancing weight to $\alpha = 0.1$; $\alpha = 0.5$ was also tested in preliminary experiments but consistently performed worse, so only $\alpha = 0.1$ is reported; this choice was made on the validation split alone. All results are reported as the mean over three random seeds: 0, 1, and 2. The implementation is available at `https://github.com/Savannah99/sf-small-bigru-kd`.

3.3 Main Comparison with SurgFormer Teacher

Compared with SurgFormer, SF-Small+BiGRU reduced the parameter count by 82.1%, from 177.8M to 31.9M, and the per-clip computation by 83.2%, from 893.28 to 150.24 GFLOPs. Both figures are measured for a single 16-frame clip at 224×224 and are dominated by the clip encoder, the BiGRU contributing a negligible share.

Table 1 shows that, on Cholec80, the CE-trained SF-Small+BiGRU remained close to the SurgFormer teacher despite an 82.1% reduction in parameter count, with Jaccard decreasing by 0.0175, from 0.7318 to 0.7143. GRU-level KD provided no further improvement, suggesting that additional soft supervision offers limited benefit in this setting. In contrast, on AutoLaparo, which contains only 10 training videos, GRU-level KD improved all metrics and increased Jaccard from 0.4997 to 0.5257. This supports the hypothesis that teacher supervision is more beneficial when training data are limited and the teacher–student gap is larger.



(a) Cholec80 (b) AutoLaparo

Fig. 2: Efficiency–performance trade-off between the SurgFormer teacher and the proposed SF-Small+BiGRU models on Cholec80 and AutoLaparo. The SF-Small+BiGRU models use approximately 82% fewer parameters while retaining Jaccard scores close to the SurgFormer teacher on Cholec80 and benefiting from GRU-level KD on AutoLaparo.

Figure 2 visualizes this trade-off: on AutoLaparo, GRU-level KD moves the student upward from the CE baseline at the same 31.9M parameter count.

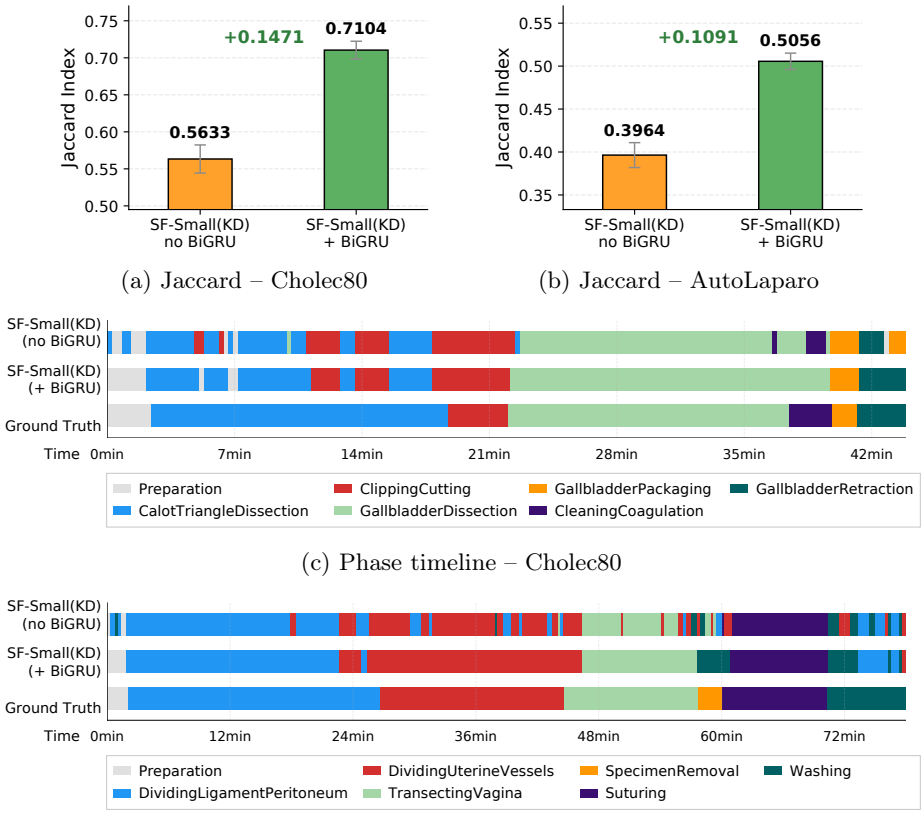
Table 1: Comparison of the SurgFormer teacher and SF-Small+BiGRU models trained with CE or GRU-level KD on Cholec80 and AutoLaparo. Student models report mean ± std over three random seeds.

Dataset	Method	Accuracy	F1	Jaccard
Cholec80	SurgFormer (Teacher)	**0.9000**	**0.8412**	**0.7318**
	SF-Small+BiGRU (CE)	0.8956±0.0053	0.8288±0.0041	0.7143±0.0052
	SF-Small+BiGRU (KD)	0.8919±0.0043	0.8271±0.0015	0.7121±0.0023
AutoLaparo	SurgFormer (Teacher)	**0.8620**	**0.7186**	**0.6245**
	SF-Small+BiGRU (CE)	0.7605±0.0197	0.6237±0.0144	0.4997±0.0182
	SF-Small+BiGRU (KD)	0.7750±0.0159	0.6433±0.0164	0.5257±0.0183

3.4 Ablation Study on BiGRU Temporal Modeling

To isolate the BiGRU contribution, we compared two SF-Small variants under the same Encoder KD setting (α=0.1, T=4.0): one predicting phases directly from encoder outputs, the other aggregating clip features with the two-layer BiGRU.

As shown in Figure 3(a,b), adding BiGRU consistently improved Jaccard on both benchmarks, from 0.5633 to 0.7104 on Cholec80 (+0.1471) and from 0.3964 to 0.5056 on AutoLaparo (+0.1091).



(a) Jaccard – Cholec80

(b) Jaccard – AutoLaparo

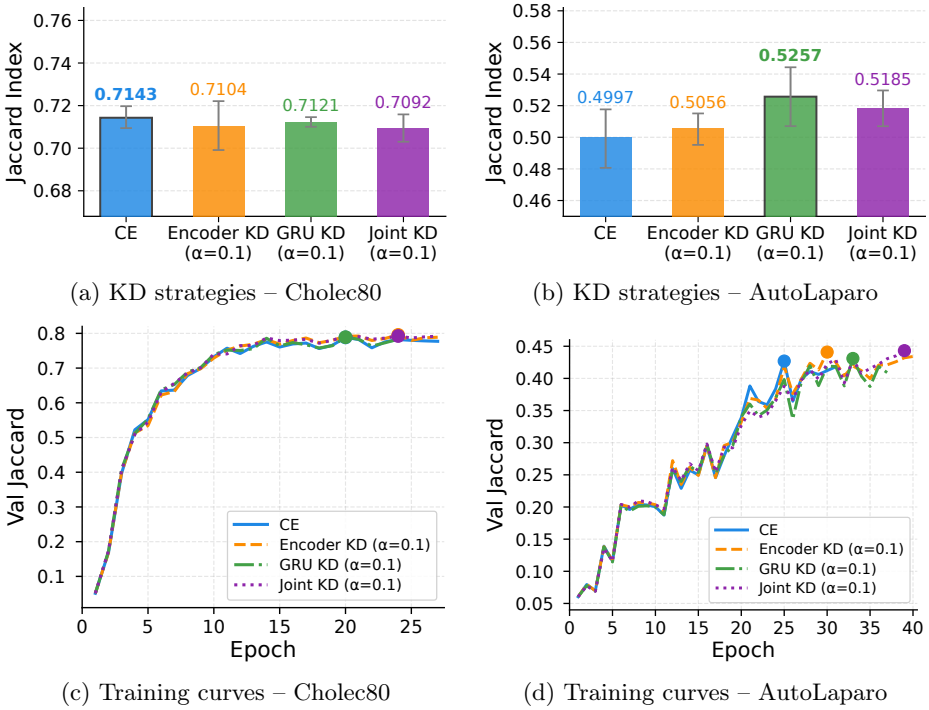
(c) Phase timeline – Cholec80

(d) Phase timeline – AutoLaparo

Fig. 3: Effect of BiGRU temporal modeling on Cholec80 and AutoLaparo. Adding BiGRU improves Jaccard on both datasets and yields smoother representative phase timelines compared with the no-BiGRU variant. Bars in (a,b) show the mean Jaccard over three seeds, with error bars indicating standard deviation. Timelines in (c,d) are visualized from the run with seed 0.

Figure 3(c,d) shows correspondingly smoother phase sequences that are better aligned with the ground truth, especially for long continuous phases. Remaining boundary errors on AutoLaparo indicate that temporal modeling remains challenging under limited training data.

Figure 4(a,b) further shows that the effect of KD is dataset-dependent. On Cholec80, all four strategies obtain similar Jaccard scores, indicating that hard-label supervision is already sufficient for the proposed framework. On AutoLaparo, GRU-level KD performs best; applying KD jointly to the encoder and the BiGRU improves over CE-only training but stays below GRU-level KD alone, so distilling into multiple branches does not necessarily yield additive gains.



(a) KD strategies – Cholec80

(b) KD strategies – AutoLaparo

(c) Training curves – Cholec80

(d) Training curves – AutoLaparo

Fig. 4: Comparison of CE-only training and three KD strategies on Cholec80 and AutoLaparo. The top row reports test Jaccard scores averaged over three seeds, with error bars indicating standard deviation. The bottom row shows validation Jaccard curves averaged over three seeds at each epoch. All KD variants use $T = 4.0$ and $\alpha = 0.1$.

Figure 4(c,d) shows similar convergence on Cholec80 and more volatile curves on AutoLaparo, likely due to the limited number of training videos. No KD strategy consistently accelerates convergence over CE-only training, suggesting that KD mainly affects final recognition performance.

Limitations. This work is a controlled teacher–student comparison and is not benchmarked against lightweight or temporal models such as SV-RCNet [8], TeCNO [2], SKiT [13], or LoViT [12], so the reported trade-off is relative to the teacher only. Efficiency is quantified by parameters and per-clip GFLOPs; video-level cost, throughput, latency, and peak memory are not measured. Sampling at 1 FPS with non-overlapping 16-frame clips yields one label per 16 seconds, quantizing boundaries and possibly favouring strong temporal smoothing; shorter clips, overlapping strides, and boundary-sensitive analysis remain future work. Finally, with three seeds the AutoLaparo KD gain is comparable to run-to-

run variation, and the bidirectional GRU needs the complete video, restricting the method to offline analysis.

4 Conclusion

We have presented a lightweight surgical phase recognition framework combining a SF-Small encoder with BiGRU-based temporal modeling. The BiGRU substantially improves clip-level predictions on both datasets, while KD gives dataset-dependent gains: limited on Cholec80, positive on AutoLaparo when applied to the temporal branch. Explicit temporal modeling should therefore be treated as a key design component, with KD placement selected empirically per dataset.

Acknowledgments. The authors would like to thank the clinical collaborators for their valuable support and discussions.

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

References

1. Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1179
2. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks. vol. 12263, pp. 343–352 (2020). https://doi.org/10.1007/978-3-030-59716-0_33
3. Demir, K.C., Schieber, H., Weise, T., Roth, D., May, M., Maier, A., Yang, S.H.: Deep Learning in Surgical Workflow Analysis: A Review of Phase and Step Recognition. IEEE Journal of Biomedical and Health Informatics **27**(11), 5405–5417 (Nov 2023). https://doi.org/10.1109/JBHI.2023.3311628
4. Ding, H., Lian, X., Unberath, M.: MoSFormer: Augmenting Temporal Context with Memory of Surgery for Surgical Phase Recognition (Mar 2025). https://doi.org/10.48550/arXiv.2503.00695
5. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer (Jul 2021). https://doi.org/10.48550/arXiv.2103.09712
6. Garrow, C.R., Kowalewski, K.F., Li, L., Wagner, M., Schmidt, M.W., Engelhardt, S., Hashimoto, D.A., Kenngott, H.G., Bodenstedt, S., Speidel, S., Müller-Stich, B.P., Nickel, F.: Machine Learning for Surgical Phase Recognition: A Systematic Review. Annals of Surgery **273**(4), 684–693 (Apr 2021). https://doi.org/10.1097/SLA.0000000000004425

7. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network (Mar 2015). https://doi.org/10.48550/arXiv.1503.02531
8. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A.: SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network. IEEE Transactions on Medical Imaging **37**(5), 1114–1126 (May 2018). https://doi.org/10.1109/TMI.2017.2787657
9. Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A.: Temporal Memory Relation Network for Workflow Recognition from Surgical Video (Mar 2021). https://doi.org/10.48550/arXiv.2103.16327
10. Junger, D., Kücherer, C., Hirt, B., Burgert, O.: Transferable situation recognition system for scenario-independent context-aware surgical assistance systems: A proof of concept. International Journal of Computer Assisted Radiology and Surgery **20**(3), 579–590 (Nov 2024). https://doi.org/10.1007/s11548-024-03283-z
11. Li, Y., Zhao, Z., Li, R., Li, F.: Deep learning for surgical workflow analysis: A survey of progresses, limitations, and trends. Artificial Intelligence Review **57**(11), 291 (Sep 2024). https://doi.org/10.1007/s10462-024-10929-6
12. Liu, Y., Boels, M., Garcia-Peraza-Herrera, L.C., Vercauteren, T., Dasgupta, P., Granados, A., Ourselin, S.: LoViT: Long Video Transformer for surgical phase recognition. Medical Image Analysis **99**, 103366 (Jan 2025). https://doi.org/10.1016/j.media.2024.103366
13. Liu, Y., Huo, J., Peng, J., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S.: SKiT: A fast key information video transformer for online surgical phase recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21074–21084 (2023). https://doi.org/10.1109/ICCV51070.2023.01927
14. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos (May 2016). https://doi.org/10.48550/arXiv.1602.03012
15. Wang, H., Jin, Y., Zhu, L.: Dynamic Interactive Relation Capturing via Scene Graph Learning for Robotic Surgical Report Generation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2702–2709 (May 2023). https://doi.org/10.1109/ICRA48891.2023.10160647
16. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: AutoLaparo: A New Dataset of Integrated Multi-tasks for Image-guided Surgical Automation in Laparoscopic Hysterectomy (Aug 2022). https://doi.org/10.48550/arXiv.2208.02049
17. Yang, S., Luo, L., Wang, Q., Chen, H.: Surgformer: Surgical Transformer with Hierarchical Temporal Attention for Surgical Phase Recognition. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, vol. 15006, pp. 606–616. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72089-5_57
18. Zhang, J., Barbarisi, S., Kadkhodamohammadi, A., Stoyanov, D., Luengo, I.: Self-Knowledge Distillation for Surgical Phase Recognition (Jun 2023). https://doi.org/10.48550/arXiv.2306.08961
19. Zou, X., Liu, W., Wang, J., Tao, R., Zheng, G.: ARST: Auto-Regressive Surgical Transformer for Phase Recognition from Laparoscopic Videos (Sep 2022). https://doi.org/10.48550/arXiv.2209.01148