RASP: Bridging the Long-Tail Gap in Surgical Video Understanding via Retrieval-Augmented Perception

Yuxiang Luo[1*], Qing Xu[2*], Yuqi Ouyang[3], and Zhen Chen[1(✉)]

[1] Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong SAR
[2] School of Computer Science, University of Nottingham, UK
[3] College of Computer Science, Sichuan University, China
z.chen@polyu.edu.hk

Abstract. Unified surgical video understanding aims to handle phase recognition, CVS assessment, and fine-grained action triplet recognition within a single framework, yet remains challenged by highly long-tailed surgical category distributions. Existing VLM-based approaches encoded category knowledge parametrically, causing rare but clinically critical triplets to be under-learned due to frequent-class dominance and fragile cross-modal alignment. To address this, we propose Retrieval-Augmented Surgical Perception (RASP), which reduces the reliance on model parameters for category knowledge by using an external memory bank that exploits standardized surgical workflows. RASP includes a Contextual Representation and Semantic Retrieval (CRSR) module that compresses temporal windows into context embeddings to recall coarse-level semantic priors and fuse them with current-frame observations. It further employs a Multi-Context Feature Representation (MCFR) module with evidence-aware cross-attention for fine-grained prior selection and an evidence gate for fallback when retrieval is uninformative. Extensive experiments show that RASP consistently outperforms state-of-the-art methods, improving AP-IVT on rare triplets by 6.3% while achieving superior multi-task performance. The code is available at https://github.com/yuxiangluo-ops/RASP.

Keywords: Surgical Video Understanding · Retrieval-Augmented Perception · Long-Tailed Recognition · Multi-Task Learning.

1 Introduction

Surgical video understanding supports intraoperative decision-making by interpreting procedural progress and clinically relevant visual cues in operative scenes [12, 11, 1, 17, 3]. Accordingly, surgical perception has been studied through multi-granularity tasks, ranging from phase recognition [7, 4] and Critical View of Safety (CVS) assessment [21] to fine-grained action triplet recognition [17].

∗ Equal contribution.

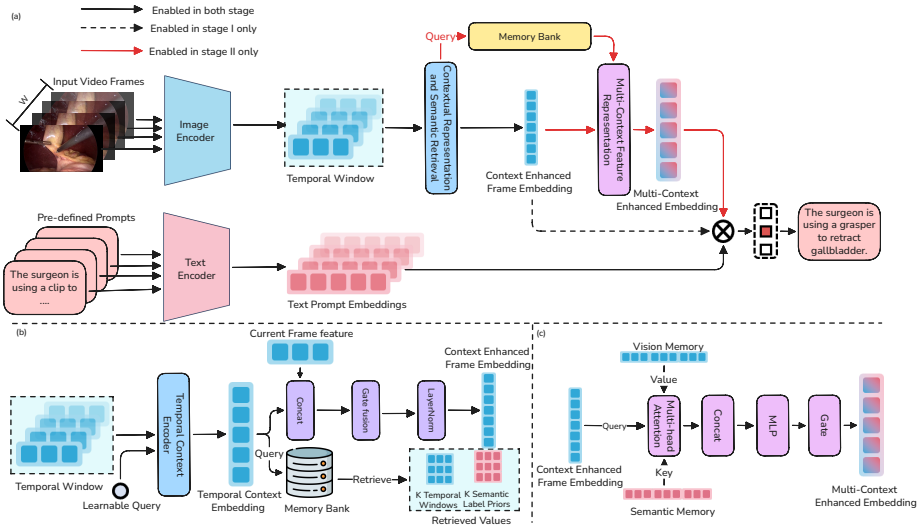

Fig. 1. Overview of the RASP framework. (a) Two-stage pipeline: Stage I (dashed line) learns context-enhanced embeddings $\mathbf{h}_t$; Stage II (red line) activates retrieval to produce $\hat{\mathbf{h}}_t$. (b) Contextual Representation and Semantic Retrieval: temporal context encoding, memory bank querying, and context-enhanced frame representation. (c) Multi-Context Feature Representation: evidence-aware cross-attention with visual key and semantic value, modulated by an evidence gate.

These heterogeneous tasks motivate unified surgical video understanding, where a single model supports multiple perception objectives.

Early efforts predominantly designed dedicated architectures for individual tasks in isolation [20, 16, 6], while recent studies have leveraged VLMs [19, 22] as a shared foundation for multi-task unification [24, 27, 23, 14]. By aligning visual representations with textual class descriptions in a shared embedding space, these VLM-based methods demonstrate that a single model can jointly handle heterogeneous surgical tasks through natural language supervision, offering a compelling and promising pathway toward surgical video understanding.

Despite recent advances, current VLM-based approaches mainly rely on parametric learning to encode category knowledge, a paradigm inherently hindered by the extremely long-tailed nature of surgical data [26, 18]. Under this imbalanced distribution, the optimization process is dominated by frequent head classes, while rare but clinically crucial triplets suffer from insufficient examples to define reliable class boundaries [5, 8]. The challenge is further exacerbated by visual-textual alignment issues, where limited visual examples for rare triplets make it difficult to form stable associations with their textual counterparts [8]. These challenges highlight the need for non-parametric retrieval as an external source of semantic priors, which can alleviate the reliance on purely parametric learning.

To address the limitations of parametric learning in surgical video understanding, we propose Retrieval-Augmented Surgical Perception (RASP), a framework that combines parametric inference with non-parametric retrieval using an external memory bank. To our best knowledge, RASP is the first framework explicitly designed to address the long-tailed distribution bottleneck in VLM-based methods for unified multi-task surgical video understanding. Our key insight leverages the fact that surgical procedures follow standardized, protocol-driven workflows [11], where recurring visually similar scenes make retrieval-augmented reasoning particularly effective. Specifically, the Contextual Representation and Semantic Retrieval (CRSR) module summarizes temporal windows into compact context embeddings, retrieves coarse-level semantic priors from the external memory bank, and integrates them with current-frame features to form context-enhanced representations. The Multi-Context Feature Representation (MCFR) module refines this process by applying evidence-aware cross-attention to select fine-grained semantic priors based on visual similarity, with an evidence gate fallback to local features when retrieval is inconclusive. Extensive experiments on CholecT50, Cholec80, and Endoscapes2023 show that RASP improves AP-IVT on rare action triplets by 6.3% and consistently outperforms baselines across multi-task evaluations.

2 Methodology

2.1 Contextual Representation and Semantic Retrieval

To reduce the reliance on model parameters for category knowledge under long-tailed surgical data, RASP stores semantic priors in an external memory bank and fuses temporal context with current-frame features to produce context-enhanced representations. As illustrated in Fig. 1, the Temporal Context Encoder produces compact temporal context embeddings as both retrieval key and inference-time query. RASP then uses these embeddings to construct the memory bank for context retrieval and fuses the temporal context embedding with the current-frame feature to obtain the context-enhanced representation.

Given a video, a frozen CLIP visual encoder extracts frame-level features $\mathbf{f}_t \in \mathbb{R}^D$, where D is the feature dimension and t indexes the current frame. We group consecutive features into an overlapping temporal window $\mathbf{T}_t = [\mathbf{f}_{t-W+1}; \ldots; \mathbf{f}_t] \in \mathbb{R}^{W \times D}$, where W is the window length and $\mathbf{T}_t$ provides the local temporal context ending at frame t.

Temporal Context Encoder. To locate historically similar surgical moments, retrieval requires representations that encode macro-level surgical progression rather than instantaneous appearance, as visually similar frames may correspond to distinct surgical states when the broader temporal context differs. We design a Temporal Context Encoder that compresses each window into a compact embedding via cross-attention pooling with a learnable query token $\mathbf{q} \in \mathbb{R}^{1 \times D}$ as

the query and $\mathbf{T}_t$ as both key and value:

$$\mathbf{z}_t = \mathrm{LN}\big(\mathrm{CrossAttn}(\mathbf{q},\ \mathbf{T}_t,\ \mathbf{T}_t)\big), \tag{1}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization. The resulting $\mathbf{z}_t \in \mathbb{R}^D$ serves as a context-level surgical state descriptor. Crucially, this embedding operates entirely within visual feature space, enabling retrieval that bypasses the fragile cross-modal alignment on which existing VLM-based methods depend.

Memory Bank Construction and Retrieval. With the temporal context encoder in place, we construct an external memory bank $\mathcal{M} = \{(\mathbf{k}_j,\ \mathbf{v}_j)\}_{j=1}^{|\mathcal{M}|}$ to store visual-semantic training entries outside model parameters. For each training window indexed by j, the key is its temporal context embedding $\mathbf{k}_j = \mathbf{z}_j$, and the value $\mathbf{v}_j = (\mathbf{T}_j,\ \mathbf{E}_j)$ pairs the frame-level visual features with corresponding semantic label embeddings $\mathbf{E}_j \in \mathbb{R}^{W\times D}$, where ground truth labels are converted to natural language descriptions and pre-encoded by the frozen CLIP text encoder [19].

At inference, the temporal context embedding $\mathbf{z}_t$ of the current window queries $\mathcal{M}$ via cosine similarity, retrieving the top-K nearest neighbors and yielding K sets of paired visual features and semantic label priors $\{(\mathbf{T}^{(k)},\ \mathbf{E}^{(k)})\}_{k=1}^{K}$. This context-level retrieval constitutes the coarse level of our hierarchical mechanism.

Context-Enhanced Frame Representation. The temporal context $\mathbf{z}_t$ captures macro-level surgical progression, whereas the current-frame feature $\mathbf{f}_t$ preserves fine-grained instantaneous observations. To adaptively weight these two sources of information at each surgical moment, we fuse them via element-wise gating as follows:

$$\mathbf{g}_t = \sigma\big(W_g\,[\mathbf{f}_t;\,\mathbf{z}_t] + \mathbf{b}_g\big), \quad \mathbf{h}_t = \mathrm{LN}\big(\mathbf{g}_t \odot \mathbf{z}_t + (1-\mathbf{g}_t)\odot \mathbf{f}_t\big), \tag{2}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes concatenation, $W_g \in \mathbb{R}^{D\times 2D}$ and $\mathbf{b}_g \in \mathbb{R}^D$ are learnable parameters, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ is the Hadamard product. The resulting $\mathbf{h}_t$ integrates temporal context with current-frame specifics, serving as the query for fine-grained aggregation of semantic priors in Section 2.2.

2.2 Multi-Context Feature Representation

The context-level retrieval in Section 2.1 recalls windows of semantically relevant priors, constituting a coarse selection based on macro-level surgical similarity. As illustrated in Fig. 1(c), given the coarse-level retrieved windows, we further perform fine-grained, frame-level aggregation of semantic priors over these candidates, using the context-enhanced representation $\mathbf{h}_t$ as the query.

We flatten the K retrieved entries into unified visual and semantic memories $\mathbf{M}_t^v, \mathbf{M}_t^e \in \mathbb{R}^{KW\times D}$. We calculate the retrieved semantic evidence $\tilde{\mathbf{h}}_t$ using the multi-head attention (MHA), as follows:

$$\tilde{\mathbf{h}}_t = \mathrm{MHA}(\mathbf{h}_t,\ \mathbf{M}_t^v,\ \mathbf{M}_t^e), \tag{3}$$

where $\mathbf{h}_t$ is the query, $\mathbf{M}_t^v$ is the key, and $\mathbf{M}_t^e$ is the value in MHA. As such, the query $\mathbf{h}_t$ and key $\mathbf{M}_t^v$ are projected into multiple parallel attention heads; within each head, attention weights are computed via scaled dot-product, and the outputs are concatenated and linearly projected back to dimension D. Critically, both the query and key are in visual feature space, while the value is textual label embeddings. The attention weights thus perform frame-level aggregation of semantic priors through visual-to-visual similarity, completing the fine-grained level of our coarse-to-fine hierarchy, while the aggregated output directly pulls the representation toward the correct semantic labels, without relying on the fragile cross-modal alignment that degrades for rare classes.

The evidence gate scalar $\lambda_t \in (0, 1)$ modulates the retrieved evidence using a gated residual, as follows:

$$\lambda_t = \sigma\big(\mathbf{w}_\lambda^\top [\mathbf{h}_t;\, \tilde{\mathbf{h}}_t] + b_\lambda\big), \quad \hat{\mathbf{h}}_t = \mathrm{LN}\big(\mathbf{h}_t + \lambda_t\, \tilde{\mathbf{h}}_t\big), \tag{4}$$

where $\mathbf{w}_\lambda \in \mathbb{R}^{2D}$ and $b_\lambda \in \mathbb{R}$ are learnable parameters. When $\lambda_t \to 0$, the model reduces to the retrieval-free baseline. The resulting $\hat{\mathbf{h}}_t$ integrates temporal context, current-frame observation, and retrieved semantic priors into a single feature for all downstream tasks.

2.3 Optimization Pipeline

The memory bank requires meaningful temporal context embeddings as the key, which are obtained only after training the Temporal Context Encoder. We therefore use a two-stage training strategy, as illustrated in Fig. 1(a). In Stage I, we train the Temporal Context Encoder and the context-enhanced frame representation. Given a temporal window $\mathbf{T}_t$, the model produces the fused representation $\mathbf{h}_t$ as described in Section 2.1. Classification is performed by computing cosine similarity between $\mathbf{h}_t$ and the text embeddings of all candidate labels, supervised by a binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\mathrm{I}} = \mathrm{BCE}\big(\cos(\mathbf{h}_t,\, \mathbf{e}_c),\; y_t^c\big), \tag{5}$$

where $\mathbf{e}_c$ is the text embedding of class c, pre-encoded from natural language descriptions by the frozen CLIP text encoder, and $y_t^c \in \{0, 1\}$ is the ground truth indicator. Upon completion, a single forward pass over the training set populates the memory bank $\mathcal{M}$.

In Stage II, we freeze all Stage I parameters and train only the evidence-aware cross-attention (Section 2.2). For each sample, we retrieve K nearest neighbors from $\mathcal{M}$ using $\mathbf{z}_t$ and produce the enhanced representation $\hat{\mathbf{h}}_t$. The same BCE loss is applied as follows:

$$\mathcal{L}_{\mathrm{II}} = \mathrm{BCE}\big(\cos(\hat{\mathbf{h}}_t,\, \mathbf{e}_c),\; y_t^c\big). \tag{6}$$

Freezing Stage I ensures that surgical context embeddings remain consistent with the key stored in $\mathcal{M}$, preventing representation drift and preserving the retrieval quality established during the first stage.

Table 1. Comparison with state-of-the-art methods on Phase Recognition, CVS Assessment, and Action Triplet Recognition. Best results are in **bold**, second best are underlined. "-" indicates that the metric was not reported in the original paper and no public checkpoint is available for reproduction.

Method	Phase	CVS Assessment				Action Triplet Recognition					
	F1	C1	C2	C3	mAP	AP-I	AP-V	AP-T	AP-IV	AP-IT	AP-IVT
Task-Specific Models											
PhaseLSTM [20]	73.1	-	-	-	-	-	-	-	-	-	-
MTRCNet [10]	77.0	-	-	-	-	-	-	-	-	-	-
SurgVLP-vision [25]	-	51.6	41.1	61.4	51.4	-	-	-	-	-	-
DeepCVS [13]	-	53.7	44.9	52.9	50.5	-	-	-	-	-	-
Rendezvous [16]	-	-	-	-	-	**92.0**	60.7	38.3	39.4	36.9	29.9
TripDis [2]	-	-	-	-	-	91.2	65.3	43.7	-	-	33.8
Multi-Task Learning Models											
ResNet50 [9]	65.1	43.2	36.9	49.8	43.3	64.0	43.0	32.7	24.4	21.9	16.7
CLIP ViT-L/14 [19]	52.6	51.4	41.2	66.0	52.9	86.5	62.4	41.5	39.3	35.2	29.3
Multi-task ResNet50 [9]	65.9	41.7	31.5	47.7	40.3	62.0	44.8	31.2	24.6	19.6	16.2
MML-SurgAdapt [21]	74.2	55.3	47.0	65.4	55.9	89.4	64.6	45.4	41.6	38.0	31.6
RASP (Ours)	**79.9**	**56.9**	**48.2**	**66.8**	**57.3**	90.5	**65.4**	**47.5**	**43.8**	**40.7**	**33.9**

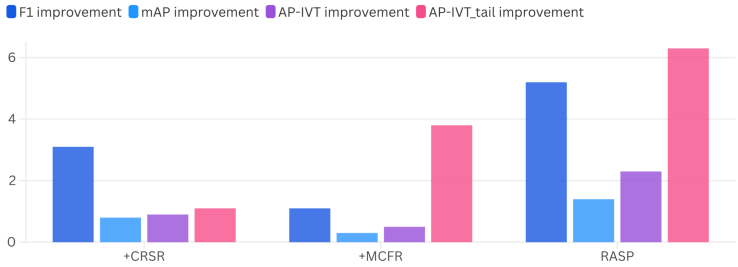

Fig. 2. Improvements of each module over the baseline on CholecT50 in action triplet recognition, highlighting the contributions of CRSR and MCFR.

3 Experiments

3.1 Experimental Setup

To validate the effectiveness of the proposed RASP, we conduct experiments on three laparoscopic cholecystectomy benchmarks: the Cholec80 dataset [20] for phase recognition, covering 7 surgical phases with a 36/7/15 train/val/test split; the CholecT50 dataset [17] for action triplet recognition, comprising 100 triplet classes with a 35/5/10 split following prior work [16]; and the Endoscapes2023 dataset [15] for CVS assessment across 3 criteria with a 120/41/40 split. Overlapping videos between the Cholec80 and CholecT50 test sets are removed to prevent data leakage. We perform all experiments on a single NVIDIA A100 GPU using PyTorch. We adopt the frozen ViT-L/14 CLIP [19] as both visual and text backbone. All frames are resized to 224×224 with CLIP normalization. Ground truth labels are converted to natural language descriptions using Gemini-3.0-Pro

Table 2. Ablation study on Cholec80, Endoscapes2023, and CholecT50. CRSR: Contextual Representation and Semantic Retrieval, MCFR: Multi-Context Feature Representation. AP-IVT$_{\text{tail}}$ denotes AP over the 25 least frequent triplet classes. †Without the Temporal Context Encoder, the memory bank is indexed by single-frame features $\mathbf{f}_t$ instead of context embeddings $\mathbf{z}_t$.

CRSR	MCFR	Phase	CVS				Triplet						
		F1	C1	C2	C3	mAP	AP-I	AP-V	AP-T	AP-IV	AP-IT	AP-IVT	AP-IVT$_{\text{tail}}$
		74.7	55.3	47.0	65.4	55.9	89.4	64.6	45.4	41.6	38.0	31.6	8.6
✓		77.8	56.1	47.8	66.2	56.7	89.8	65.4	46.2	42.3	38.7	32.5	9.7
	✓†	75.8	55.6	47.3	65.7	56.2	89.8	64.9	45.9	42.0	38.3	32.1	12.4
✓	✓	**79.9**	**56.9**	**48.2**	**66.8**	**57.3**	**90.5**	**65.4**	**47.5**	**43.8**	**40.7**	**33.9**	**14.9**

and pre-encoded by the frozen CLIP text encoder. Separate memory banks are constructed per dataset. The temporal window size W and neighbor count K are set to 6 and 3, respectively. Both stages use the Adam optimizer with a learning rate of 1×10^{-5} and a batch size of 128, each training for 15 epochs. The memory bank is built via a single forward pass after Stage I, and all Stage I parameters are frozen during Stage II. For quantitative evaluation, we report F1-score for phase recognition, frame-level AP for each CVS criterion (C1–C3) and mAP for CVS assessment, and mAP for instrument (I), verb (V), target (T), and their compositions (IV, IT, IVT) in triplet recognition, following established protocols [20, 16, 15].

3.2 Comparison with State-of-the-art Methods

As shown in Table 1, we compare RASP with both task-specific and multi-task methods. Task-specific approaches [25, 2] achieve competitive results on their respective objectives, *e.g.*, TripDis attains 33.8% AP-IVT and SurgVLP-vision reaches 51.4% CVS mAP, yet both methods remain limited to their respective target tasks. Among multi-task models, MML-SurgAdapt [21] is the strongest parametric baseline, achieving 74.2% Phase F1, 55.9% CVS mAP, and 31.6% AP-IVT. RASP further improves upon MML-SurgAdapt, reaching 79.9% Phase F1, 57.3% CVS mAP, and 33.9% AP-IVT, and achieves the best multi-task performance across the reported metrics. Notably, RASP also exceeds the best task-specific methods despite being a unified model, outperforming TripDis on triplet recognition and SurgVLP-vision on CVS mAP by 5.9%. These results suggest that retrieval augmentation provides useful semantic priors for unified surgical video understanding.

3.3 Ablation Study

To validate the effectiveness of the proposed CRSR and MCFR modules, we conduct ablation studies on Cholec80, Endoscapes2023, and CholecT50, as shown in Table 2. By introducing CRSR alone, the model achieves 3.1% Phase F1 and 0.9% AP-IVT improvements, confirming that temporal context provides

Table 3. Effect of window size W.

W	Phase (F1)	CVS (mAP)	AP-IVT
2	76.5	56.4	32.3
4	78.2	56.8	33.1
6	79.9	**57.3**	**33.9**
8	**80.5**	57.1	33.6

Table 4. Effect of neighbor count K.

K	Phase (F1)	CVS (mAP)	AP-IVT
1	78.4	56.8	32.9
2	79.2	57.1	33.5
3	**79.9**	**57.3**	**33.9**
4	79.6	57.4	33.4

discriminative information beyond single-frame appearance. Since MCFR relies on context embeddings from CRSR for memory bank indexing, we isolate it by substituting single-frame features $\mathbf{f}_t$ as the retrieval key (marked as † in Table 2), which yields only marginal gains of 0.5% AP-IVT, validating that frame-level features can hardly serve as reliable retrieval indices. Notably, incorporating both modules brings the largest boost across all metrics, improving AP-IVT by 2.3% and AP-IVT$_{\text{tail}}$ by 6.3% over the baseline. This result shows that retrieval augmentation particularly benefits rare, clinically critical triplets. As visualized in Fig. 2, the gains from CRSR and MCFR are largely complementary, with AP-IVT$_{\text{tail}}$ benefiting the most. These comprehensive results validate that the tailored CRSR and MCFR collectively contribute to the superior performance of RASP across diverse surgical tasks.

3.4 Hyperparameter Analysis

We further investigate the sensitivity of RASP to two key hyperparameters: the temporal window length W and the number of retrieved neighbors K. As shown in Table 3, performance improves consistently as W increases from 2 to 6, confirming that broader temporal context yields more informative context embeddings. At W=8, phase recognition continues to improve to 80.5% F1, but triplet detection slightly degrades to 33.6% AP-IVT, suggesting that excessively long windows dilute retrieval specificity for fine-grained recognition. We therefore select W=6 as the default. As shown in Table 4, performance peaks at K=3 with 33.9% AP-IVT and 57.3% CVS mAP, while K=4 shows a slight decline, indicating that a small set of high-quality retrievals suffices while additional neighbors introduce noise. The modest gap of 1.0% AP-IVT between K=1 and K=3 further shows that the fusion module effectively aggregates complementary evidence rather than relying on a single best match. Overall, RASP consistently outperforms the parametric baseline across all tested configurations of W and K, demonstrating robustness to hyperparameter choices.

4 Conclusion

In this work, we present RASP, a retrieval-augmented framework for unified surgical video understanding that addresses the long-tail bottleneck inherent in purely parametric VLM-based approaches. We devise a CRSR module that compresses temporal windows into context embeddings to query an external

memory bank, decoupling tail-class preservation from gradient-driven training dynamics. To effectively exploit such coarse-level retrieved semantic priors, we propose a MCFR module that performs evidence-aware cross-attention for fine-grained aggregation of semantic priors, modulated by an evidence gate that enables fallback to local features when retrieval is uninformative. Extensive experiments on CholecT50, Cholec80, and Endoscapes2023 demonstrate that RASP consistently outperforms both task-specific and multi-task baselines across all metrics, including a 6.3% gain in AP-IVT on rare triplets, supporting its effectiveness for unified surgical video understanding.

Acknowledgments. This work was supported by PolyU Start-up Fund P0060371.

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

References

1. Alabi, O., Vercauteren, T., Shi, M.: Multitask learning in minimally invasive surgical vision: A review. Med. Image Anal. **101**, 103480 (2025)
2. Chen, Y., He, S., Jin, Y., Qin, J.: Surgical activity triplet recognition via triplet disentanglement. In: MICCAI. pp. 451–461. Springer (2023)
3. Chen, Z., Guo, Q., Yeung, L.K., Chan, D.T., Lei, Z., Liu, H., Wang, J.: Surgical video captioning with mutual-modal concept alignment. In: MICCAI. pp. 24–34. Springer (2023)
4. Chen, Z., Zhai, Y., Zhang, J., Wang, J.: Surgical temporal action-aware network with sequence regularization for phase recognition. In: BIBM. pp. 1836–1841. IEEE (2023)
5. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR. pp. 9268–9277 (2019)
6. 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. In: MICCAI. pp. 343–352. Springer (2020)
7. Gao, X., Jin, Y., Long, Y.L., Dou, Q., Heng, P.A.: Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: MICCAI. pp. 593–603. Springer (2021)
8. Gui, S., Wang, Z.: Tail-enhanced representation learning for surgical triplet recognition. In: MICCAI. pp. 689–699. Springer (2024)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
10. Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J.Q., Fu, C.W., Heng, P.A.: Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med. Image Anal. **59**, 101572 (2020)
11. Lalys, F., Jannin, P.: Surgical process modelling: a review. Int. J. Comput. Assist. Radiol. Surg. **9**(3), 495–511 (2014)
12. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al.: Surgical data science for next-generation interventions. Nat. Biomed. Eng. **1**(9), 691–696 (2017)

13. Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., Fiorillo, C., Pessaux, P., Mutter, D., Marescaux, J., Costamagna, G., et al.: Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Ann. Surg. **275**(5), 955–961 (2022)
14. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
15. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., et al.: The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. arXiv preprint arXiv:2312.12429 (2023)
16. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Med. Image Anal. **78**, 102433 (2022)
17. Nwoye, C.I., Yu, T., Sharma, S., Murali, A., Alapatt, D., Vardazaryan, A., Yuan, K., Hajek, J., Reiter, W., Yamlahi, A., et al.: Cholectriplet2022: Show me a tool and tell me the triplet—an endoscopic vision challenge for surgical action triplet detection. Med. Image Anal. **89**, 102888 (2023)
18. Pan, L., Zhang, Y., Yang, Q., Li, T., Chen, Z.: Combat long-tails in medical classification with relation-aware consistency and virtual features compensation. In: MICCAI. pp. 14–23. Springer (2023)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
20. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging **36**(1), 86–97 (2016)
21. Walimbe, S., Baby, B., Srivastav, V., Padoy, N.: Adaptation of multi-modal representation models for multi-task surgical computer vision. In: MICCAI. pp. 24–33. Springer (2025)
22. Wang, G., Bai, L., Wang, J., Yuan, K., Li, Z., Jiang, T., He, X., Wu, J., Chen, Z., Lei, Z., et al.: Endochat: Grounded multimodal large language model for endoscopic surgery. Med. Image Anal. p. 103789 (2025)
23. Yuan, K., Navab, N., Padoy, N., et al.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. NeurIPS **37**, 122952–122983 (2024)
24. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: MICCAI. pp. 306–316. Springer (2024)
25. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Marescaux, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. Med. Image Anal. **105**, 103644 (2025)
26. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(9), 10795–10816 (2023)
27. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vis. **130**(9), 2337–2348 (2022)