SemiVT-Surge: Semi-Supervised Video Transformer for Surgical Phase Recognition

Yiping Li[1], Ronald de Jong[1], Sahar Nasirihaghighi[2], Tim Jaspers[3], Romy van Jaarsveld[4], Gino Kuiper[4], Richard van Hillegersberg[4], Fons van der Sommen[3], Jelle Ruurda[4], Marcel Breeuwer[1], and Yasmina Al Khalil[1]

[1] Department of Biomedical Engineering, Medical Image Analysis, Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Institute of Information Technology (ITEC), University of Klagenfurt, Austria
[3] Department of Electrical Engineering, Video Coding & Architectures, Eindhoven University of Technology, Eindhoven, The Netherlands
[4] Department of Surgery, University Medical Center Utrecht, Utrecht, The Netherlands

Abstract. Accurate surgical phase recognition is crucial for computer-assisted interventions and surgical video analysis. Annotating long surgical videos is labor-intensive, driving research toward leveraging unlabeled data for strong performance with minimal annotations. Although self-supervised learning has gained popularity by enabling large-scale pretraining followed by fine-tuning on small labeled subsets, semi-supervised approaches remain largely underexplored in the surgical domain. In this work, we propose a video transformer-based model with a robust pseudo-labeling framework. Our method incorporates temporal consistency regularization for unlabeled data and contrastive learning with class prototypes, which leverages both labeled data and pseudo-labels to refine the feature space. Through extensive experiments on the private RAMIE (Robot-Assisted Minimally Invasive Esophagectomy) dataset and the public Cholec80 dataset, we demonstrate the effectiveness of our approach. By incorporating unlabeled data, we achieve state-of-the-art performance on RAMIE with a 4.9% accuracy increase and obtain comparable results to full supervision while using only 1/4 of the labeled data on Cholec80. Our findings establish a strong benchmark for semi-supervised surgical phase recognition, paving the way for future research in this domain. Code is available at https://github.com/IntraSurge/SemiVT-Surge.

Keywords: Surgical Phase Recognition · Semi-supervised Learning · Video Transformer

1 Introduction

Surgical phase recognition is used in computer-assisted surgery systems to classify each frame from surgical video footage into different stages of a surgical procedure. It supports context-aware assistance and decision support, enhances workflow efficiency, facilitates postoperative analysis, surgeon performance evaluation, and identification of problematic phases [17]. It remains a high-level understanding task, presenting significant challenges as the model must distinguish between similar semantics across different phases of a procedure.

Transformers have become increasingly popular in surgical phase recognition, with tailored attention mechanisms to enhance spatio-temporal modeling. SKiT [16] introduced critical pooling to record key information and self- and cross-attention for feature fusion. Label-Guided Teacher [10] incorporated labels as extra supervision via cross-attention to refine feature representations. To better capture temporal dynamics of video data, researchers have recently shifted to video-based models, instead of relying on per-frame feature extraction. Surgformer [26] enhanced TimeSformer [5] with Hierarchical Temporal Attention to improve spatio-temporal feature capture, while MuST [18] proposed a multi-scale Transformer that captures short-, mid-, and long-term dependencies.

Annotating surgical phases is labor-intensive, requiring a surgical expert to watch the full procedure and annotate key intervals. To reduce supervision, self-supervised learning has gained traction, leveraging large surgical datasets for pretraining before fine-tuning on downstream tasks such as phase recognition and semantic segmentation [1,4,12,13] , demonstrating promising results. Surgery-specific pretraining has been shown to improve performance, whereas cross-surgery pretraining may even underperform compared to models trained on general computer vision datasets [1]. Given the procedure-specific nature of surgical phases, semi-supervised learning offers a promising way to leverage unlabeled data from the same surgery and reduce annotation effort. However, it remains underexplored in surgical phase recognition, with only a few studies addressing its potential. SurgSSL [19] introduced a two-stage semi-supervised approach with visual and temporal consistency as regularization and pseudo-labeling for refinement, while FedCy [14] applied semi-supervised learning in a federated setting using temporal cycle consistency and contrastive learning for multi-center data. Both methods, developed early on, were based on simple convolutional neural networks.

Beyond surgical applications, semi-supervised learning has been widely explored in image and short-term video tasks. FixMatch [20] introduced consistency regularization with weak and strong augmentations for image classification, while Mean Teacher [21] demonstrated the effectiveness of an exponential moving average (EMA) strategy for semi-supervised learning. More recent advances [8,24,25] have further refined augmentation strategies and temporal modeling for short-term action recognition tasks, yet their applicability to long-duration tasks like surgical phase recognition remains unexplored. Contrastive learning is well-established for learning discriminative representations. In supervised settings, it has proven effective for surgical phase recognition [10]. In semi-supervised medical image classification and segmentation, class prototypes have been used to structure the feature space around class-specific embeddings [3,11], improving learning with pseudo-label guidance, which inspired our work.

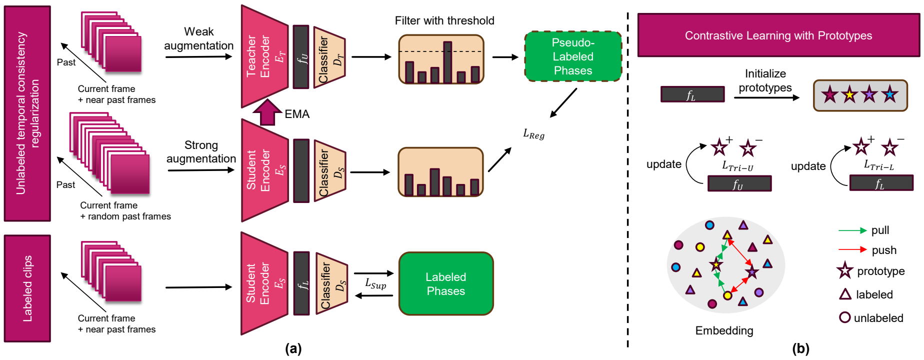

Fig. 1. Overview of the proposed method: (a) overall training process with temporal consistency regularization, as detailed in Section 2.2, and (b) contrastive learning with class prototypes, as detailed in Section 2.3.

In this work, we explore the potential of semi-supervised learning for surgical phase recognition using a video transformer model. An overview of our approach is provided in Fig. 1. The training procedure involves two key components: (a) long- and short-term frame sampling for weak and strong augmentations during pseudo-labeling to improve temporal consistency; and (b) contrastive learning with class prototypes to regularize the feature space using both labeled and pseudo-labeled data, with prototypes continuously updated to refine feature embeddings. We demonstrate the superiority of our method over state-of-the-art approaches through comprehensive experiments on two datasets.

2 Methods

2.1 Overview of the training strategy

In this work, we address the problem of online surgical phase recognition, a video classification task that predicts the surgical phase Y_t of the current frame from video input I_t. Given a video stream $I_t = \{I_j\}_{j=1}^{t}$ up to time t, we aim to learn a mapping M_θ such that $M_\theta(I_t) \approx Y_t$, where $Y_t \in \{1, \ldots, C\}$ represents the surgical phase at time t and C is the total number of phases. We use TimeSformer [5] as the video encoder E, which extracts spatiotemporal features $f_t = E(I_t)$. Here, f_t is the class token with a feature embedding size of 768. The input consists of $T = 16$ frames, including the current frame and 15 past frames. A classifier D, implemented as a linear layer, predicts the phase probability distribution $Y_t = D(f_t)$. θ denotes model parameters. The success of this task relies on training a robust encoder E for high-quality feature representations. The training begins with a warm-up on labeled data using a standard cross-entropy loss L_{Sup}, followed by semi-supervised training with both labeled and unlabeled data, where the supervised loss L_{Sup} is still applied to the labeled data, while additional losses are introduced. Specifically, the temporal consistency loss L_{Reg} is applied solely to the unlabeled data (see Section 2.2 for details). The contrastive losses $L_{\text{Tri-U}}$ and $L_{\text{Tri-L}}$ regularize the feature embeddings using updated class prototypes with both labeled and unlabeled data (see Section 2.3 for details). The total loss L_{total} is the sum of these individual losses. For a detailed description of the full workflow, refer to Algorithm 1.

Algorithm 1 Workflow of Our Proposed Method

```
Require: Labeled dataset $\mathcal{S}$, unlabeled dataset $\mathcal{U}$
  Warm-up: Train $E_S, D_S$ with $L_{\text{Sup}}$ to initialize $\theta_{E,S}, \theta_{D,S}$.
  Teacher Init: Set $\theta_{E,T} \leftarrow \theta_{E,S}, \theta_{D,T} \leftarrow \theta_{D,S}$.
  Prototypes Initialization: mean features from labeled dataset: $\mathcal{C} \leftarrow \{\mu_c\}_{c=1}^{C}$
  where $\mu_c = \frac{1}{|I_L^c|} \sum Normalize(E_S(I_L^c))$
  while iteration $\leq$ max iteration do
      Sample mini-batches $B_L$ from $\mathcal{S}$ and $B_U$ from $\mathcal{U}$
      for $(I_L, Y_L) \in B_L$ do
          $Y'_L \leftarrow D_S(E_S(I_L))$                                    ▷ Student prediction
          $L_{\text{Sup}} \leftarrow \text{CE}(Y'_L, Y_L)$                         ▷ Supervised loss
          $f_L \leftarrow \text{Normalize}(E_S(I_L))$                          ▷ Feature embedding
          $L_{\text{Tri-L}} \leftarrow \text{TripletLoss}(f_L, \mathcal{C}_{Y_L}, \{\mathcal{C}_k\}_{k \neq Y_L})$
          Update prototypes: $\mathcal{C}_{Y_L} \leftarrow \eta \mathcal{C}_{Y_L} + (1 - \eta) f_L$
      end for
      for $I_U \in B_U$ do
          $I_U^s \leftarrow \text{StrongAugment}(I_U)$                      ▷ Strong augmentation
          $I_U^w \leftarrow \text{WeakAugment}(I_U)$                          ▷ Weak augmentation
          $Y'_{U,S} \leftarrow D_S(E_S(I_U^s))$                               ▷ Student prediction
          $Y'_{U,T} \leftarrow D_T(E_T(I_U^w))$                               ▷ Teacher prediction
          $f_U \leftarrow \text{Normalize}(E_T(I_U^w))$                       ▷ Feature embedding
          Filter $\mathcal{F} \leftarrow \{(f_U, Y'_{U,S}, Y'_{U,T}) | \max(Y'_{U,T}) \geq \delta\}$  ▷ High-confidence subset only
          for $(f_U, Y_S, Y_T) \in \mathcal{F}$ do
              $L_{\text{Tri-U}} \leftarrow \text{TripletLoss}(f_U, \mathcal{C}_{\arg\max Y_T}, \{\mathcal{C}_k\}_{k \neq \arg\max Y_T})$
              $L_{\text{Reg}} \leftarrow \text{CE}(Y_S, Y_T)$                  ▷ Temporal Consistency Regularization
              Update prototypes: $\mathcal{C}_{Y_T} \leftarrow \eta \mathcal{C}_{Y_T} + (1 - \eta) f_U$
          end for
      end for
      Aggregate losses: $L_{\text{total}} \leftarrow L_{\text{Sup}} + L_{\text{Reg}} + L_{\text{Tri-U}} + L_{\text{Tri-L}}$
      Update student: $\theta_{E,S}, \theta_{D,S} \leftarrow \theta - \nabla_\theta L_{\text{total}}$
      Update teacher: $\theta_{E,T}, \theta_{D,T} \leftarrow \alpha \theta_T + (1 - \alpha) \theta_S$            ▷ EMA
  end while
  return $\theta_{E,T}, \theta_{D,T}$
```

2.2 Temporal Consistency Regularization

Our semi-supervised learning approach enforces consistency between differently augmented views of unlabeled data using a teacher-student architecture. The student model is shared across both labeled and unlabeled batches, and during training, the student model receives updates via gradients, while the teacher model is updated using EMA. Considering the nature of surgical phase recognition, temporal sampling itself serves as a natural augmentation strategy, enabling the model to learn robust representations across different temporal contexts. Given an unlabeled batch B_U, each sample $I_U \in B_U$ undergoes two augmentations: a weakly augmented view $I_U^w = \text{WeakAugment}(I_U)$, which preserves short-term temporal context by including the processing frame and its preceding $T-1$ consecutive frames, and a strongly augmented view $I_U^s = \text{StrongAugment}(I_U)$, which enforces long-term temporal alignment by randomly selecting $T-1$ frames from the full video history. Further details on the augmentation process can be found in Section 2.4. The student model predicts $Y'_{U,S} = D_S(E_S(I_U^s))$, while the teacher model, predicts $Y'_{U,T} = D_T(E_T(I_U^w))$. To ensure reliable pseudo-labeling, we introduce a confidence-based selection mechanism: only samples where the maximum predicted class probability exceeds a predefined threshold δ contribute to the unsupervised loss. Formally, the loss is defined as

$$\mathcal{L}_U = \frac{1}{N_U} \sum_{I_U \in B_U} \mathbb{I}\left(\max(Y'_{U,T}) > \delta\right) \text{CE}(Y'_{U,S}, Y'_{U,T}), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 when the confidence threshold is met, filtering out uncertain pseudo-labels, and CE denotes the standard cross-entropy loss. This dual-stream approach, coupled with confidence-based selection, facilitates effective feature learning while reducing the influence of noisy pseudo-labels.

2.3 Contrastive Learning with Prototypes

Our method integrates semi-supervised learning with contrastive prototype alignment to regularize the feature embedding space. We employ triplet margin loss, computed in Euclidean space, to structure the embedding space by enforcing a margin between positive and negative samples. Normalizing the features ensures that distance measurements remain consistent and robust, thereby enhancing class separability. Class-specific prototypes are initialized as cluster centroids in a normalized feature space computed from labeled data with warmed-up model. These prototypes evolve through EMA updates during training. Unlabeled samples contribute only when teacher-generated pseudo-labels exceed a confidence threshold δ, filtering out unreliable labels as described in Equation 1.

The prototype-anchored triplet loss enforces discriminative feature learning by pulling samples toward their class prototype while repelling them from negative prototypes:

$$\mathcal{L}_{\text{tri}} = \max\left(d(f, \mathcal{C}_y) - d\left(f, \frac{1}{k}\sum_{i=1}^{k} \mathcal{C}_{\text{neg},i}\right) + m, 0\right), \tag{2}$$

where $d(\cdot,\cdot)$ denotes the Euclidean distance, m is the margin (empirically set to 0.3), and the negative prototype $\mathcal{C}_{\text{neg}}$ is computed as the average of the k hardest negative prototypes, with $k = 3$ selected as the nearest negative prototypes to the feature f. For labeled data, positive and negative prototypes are selected based on the ground truth labels, while for unlabeled data, they are determined according to the pseudo-labels generated by the teacher model Y_T.

2.4 Implementation details

We use the TimeSformer [5] model, initialized with pre-trained weights from the Kinetics-400 dataset [15], which has been shown to accelerate convergence. Video frames are pre-extracted at 1fps, and resized to 256 × 256. We first warm up with training only on labeled data for 3 epochs, and perform semi-supervised training for another 12 epochs. Data augmentation is performed using the AutoAugment library [6], with a strong augmentation configuration of 'rand-m9-n5-mstd0.8-inc1'. For weak augmentation, we apply random cropping, normalization, random rotation, and random crop. All models are trained with a batch size 16 on a single NVIDIA H100 GPU, with a SGD optimizer using a momentum of 0.9 and a weight decay of 0.001. For each setting, the basic learning rate is set to 0.005, and halved at epoch 8 and 12. The confidence threshold δ was emperically chosen as 0.6 for RAMIE dataset and 0.8 for Cholec80. The EMA update rules are detailed in Algorithm 1, with the EMA parameters for updating both the teacher model and prototypes set to $\alpha = \eta = 0.9$.

3 Experiments and results

Datasets. Cholec80 [22]: This dataset consists of 80 cholecystectomy surgery videos, each annotated with 7 distinct surgical phases. The dataset is divided into official training and test sets, with each containing 40 videos. For experiments with reduced annotated data, we randomly sample a subset of the training set as labeled data, while treating the remaining videos as unlabeled.

RAMIE [2]: The RAMIE dataset is a private dataset consisting of 27 labeled Robot-Assisted Minimally Invasive Esophagectomy videos. The dataset is split into 14 videos for training, 4 for validation, and 9 for testing, with annotations for 13 surgical phases. Compared to cholecystectomy, RAMIE surgery exhibits more complex temporal dynamics, with phase repetitions and increased variability in phase sequences. The unlabeled data is drawn from the same collection, with 20 videos randomly selected from a pool of 70 available unlabeled videos.

Evaluation details and metrics. Our model is evaluated in an online manner, sliding through the video with a window size of $T = 16$ to generate frame-wise predictions. The evaluation metrics include widely-used benchmark metrics: video-level Accuracy, phase-level Precision, phase-level Recall, phase-level Jaccard, and phase-level F1 score, implemented with the code from [9]. Mean ± standard deviation is computed across videos in the test set.

Ablation studies. We conduct ablation studies to assess the effectiveness of each component during training. Among these, the results for supervised learning and Temporal Consistency Regularization (TCR) can be compared to the

Table 1. Ablative testing results on RAMIE and Cholec80 datasets. Sup. indicates supervised training with labeled data only. TCR represents the use of Temporal Consistency Regularization, CLP represents the inclusion of Contrastive Learning with Prototypes, and TCN represents the addition of TeCNO's causal TCN. A checkmark (✓) indicates the presence of an attribute, while (-) denotes its absence.

Sup.	TCR	CLP	TCN	RAMIE				Cholec80 (20 labeled training)			
				Accuracy	Precision	Recall	Jaccard	Accuracy	Precision	Recall	Jaccard
✓	-	-	-	78.7 ± 4.2	79.7 ± 4.0	74.7 ± 5.2	60.9 ± 4.3	84.3 ± 10.7	81.4 ± 9.0	79.0 ± 9.4	65.8 ± 13.2
✓	✓	-	-	80.4 ± 3.3	79.8 ± 3.4	75.6 ± 4.9	62.2 ± 3.9	87.0 ± 6.6	83.7 ± 8.3	82.9 ± 6.8	70.7 ± 10.2
✓	✓	✓	-	81.9 ± 2.9	80.6 ± 3.8	78.9 ± 3.7	64.9 ± 3.4	89.6 ± 7.0	86.5 ± 8.0	86.5 ± 6.3	75.9 ± 10.7
✓	✓	✓	✓	**83.4 ± 3.7**	**81.6 ± 3.4**	**78.8 ± 3.8**	**66.1 ± 3.1**	**90.4 ± 7.0**	**88.4 ± 7.2**	**86.5 ± 7.4**	**77.5 ± 11.3**

popular semi-supervised learning method, FixMatch [20], which highlights the effectiveness of our temporal regularization.

With the addition of Contrastive Learning with Prototypes (CLP), we observe a more significant improvement. Since surgical phase recognition is essentially a classification task, introducing greater contrast between classes proves to be highly beneficial. This serves as a regularization method within the embedding space.

Our video transformer model can function independently by processing the current frame along with past frames, inherently capturing temporal dynamics. TeCNO [7], widely used as a classification head in surgical phase recognition, is particularly common in self-supervised model evaluations. To assess its impact, we integrate TeCNO's causal TCN, where our model functions as a spatio-temporal feature extractor. This distinguishes it from approaches that rely solely on frame-wise feature extraction. The results demonstrate that TeCNO further improves performance, as it excels at capturing long-term temporal dependencies. Results are shown in Table 1 and qualitative results and F1 across phases on both datasets can be found in Figure 2 and Figure 3.

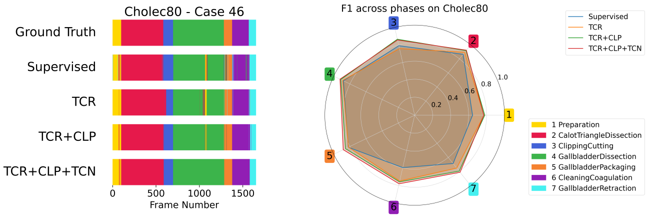

Fig. 2. Cholec80: Qualitative result (left) and mean F1 scores across phases (right).

State-of-the-art Comparison. We compare our method to state-of-the-art approaches on both datasets. Given the limited research on semi-supervised learning for surgical phase recognition, we also evaluate self-supervised methods, including self-pretrained models with a temporal module based on TeCNO

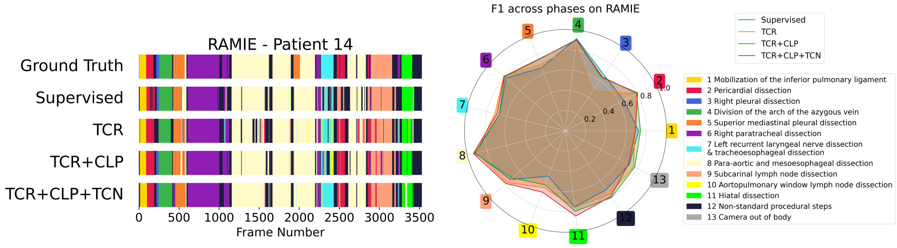

Fig. 3. RAMIE: Qualitative result (left) and mean F1 scores across phases (right).

[7]. The results are shown in Table 2. EndoFM [23] was trained using DINO on over 33,000 endoscopic video clips, and SurgeNetXL [13] was trained on 4,711,024 frames from 23 surgical procedures. Compared to these two large-scale self-pretraining approaches, our model outperforms both, achieving a 3% and 0.9% accuracy increase, respectively, while using significantly less surgery-specific unlabeled data. We also explored the use of additional unlabeled videos (beyond 20) on the RAMIE dataset, but did not observe further improvements in performance, which warrants further investigation. On the Cholec80 dataset, [1] reported an F1 score of 67.4 ± 4.9 using 5 labeled videos for training. Our method achieves comparable results. Endovit [4] reported an accuracy of 84.68 ± 1.25 using 8 labeled videos, while our method performed similarly with only 5 labeled videos. Since these papers did not provide additional evaluation metrics, they are excluded from Table 3. As shown in Table 3, our method outperforms the semi-supervised SurgSSL [19] and matches the full supervision TeCNO performance on Cholec80, using only 1/4 of the labeled data.

Table 2. Comparison with state-of-the-art methods on RAMIE dataset.

Method	Supervision Type	Accuracy (%)	Precision (%)	Recall (%)	Jaccard (%)
TeCNO [7]	Fully supervised	78.5 ± 4.0	73.9 ± 4.6	73.6 ± 5.1	58.3 ± 4.8
FixMatch [20]	Semi-supervised	79.2 ± 4.1	80.0 ± 4.1	75.5 ± 5.6	61.8 ± 4.8
EndoFM [23]	Self-supervised pretraining	80.4 ± 3.3	79.0 ± 2.9	74.6 ± 5.6	61.9 ± 4.7
SurgeNetXL [13]	Self-supervised pretraining	82.5 ± 3.7	79.4 ± 3.9	78.7 ± 3.9	64.8 ± 4.1
Ours	Semi-supervised	**83.4 ± 3.7**	**81.6 ± 3.4**	**78.8 ± 3.8**	**66.1 ± 3.1**

4 Conclusion

In this work, we propose a semi-supervised framework for surgical phase recognition with a video transformer model. By incorporating long-short term temporal sampling and dynamic contrastive learning with class prototypes, our method effectively leverages unlabeled videos, demonstrating strong performance on both the Cholec80 and RAMIE datasets. Future work can explore better pseudo-label

Table 3. Comparison with state-of-the-art methods on the Cholec80 dataset.

Method	Supervision Type	Labeled Videos	Accuracy (%)	Precision (%)	Recall (%)	Jaccard (%)
TeCNO [7]	Fully supervised	40	88.6 ± 7.8	86.5 ± 7.0	87.6 ± 6.7	75.1 ± 6.9
SurgSSL [19]	Semi-supervised	20	87.0 ± 7.4	84.2 ± 8.9	85.2 ± 11.1	70.5 ± 12.6
Ours			**90.4 ± 7.0**	**88.4 ± 7.2**	**86.5 ± 7.4**	**77.5 ± 11.3**
SurgSSL [19]	Semi-supervised	10	85.0 ± 7.7	83.3 ± 8.3	83.1 ± 12.3	68.0 ± 13.5
Ours			**88.8 ± 6.5**	**85.0 ± 7.8**	**86.2 ± 6.1**	**74.5 ± 10.4**
SurgSSL [19]	Semi-supervised	5	83.2 ± 7.7	81.8 ± 9.9	81.6 ± 12.3	65.6 ± 14.8
Ours			**85.3 ± 7.6**	**84.2 ± 7.6**	**81.5 ± 7.8**	**69.2 ± 10.3**

filtering to enhance performance on underrepresented classes, evaluate its effectiveness in federated learning, and analyze clinical relevance and failure cases for further improvements.

Acknowledgments. This research was funded by Stichting Hanarth Fonds, study number: 2022-13. It is part of the INTRA-SURGE (INTelligent computeR-Aided Surgical gUidance for Robot-assisted surGEry) project aimed at advancing the future of surgery.

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

References

1. Alapatt, D., Murali, A., Srivastav, V., Consortium, A., Mascagni, P., Padoy, N.: Jumpstarting surgical computer vision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 328–338. Springer (2024)
2. Author, A.: Paper title. In: This paper is accepted by XXX but not yet available to the public. More details on the dataset are provided in this work. (2025)
3. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19786–19797 (2023)
4. Batić, D., Holm, F., Özsoy, E., Czempiel, T., Navab, N.: Endovit: pretraining vision transformers on a large collection of endoscopic images. International Journal of Computer Assisted Radiology and Surgery **19**(6), 1085–1091 (2024)
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
6. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 113–123 (2019)
7. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 343–352. Springer (2020)

8. Dave, I.R., Rizve, M.N., Chen, C., Shah, M.: Timebalance: Temporally-invariant and temporally-distinctive video representations for semi-supervised action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2341–2352 (2023)
9. Funke, I., Rivoir, D., Speidel, S.: Metrics matter in surgical phase recognition. arXiv preprint arXiv:2305.13961 (2023)
10. Guan, J., Zou, X., Tao, R., Zheng, G.: Label-guided teacher for surgical phase recognition via knowledge distillation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 349–358. Springer (2024)
11. He, A., Li, T., Zhao, Y., Zhao, J., Fu, H.: Open-set semi-supervised medical image classification with learnable prototypes and outlier filter. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 492–501. Springer (2024)
12. Jaspers, T.J., de Jong, R.L., Al Khalil, Y., Zeelenberg, T., Kusters, C.H., Li, Y., van Jaarsveld, R.C., Bakker, F.H., Ruurda, J.P., Brinkman, W.M., et al.: Exploring the effect of dataset diversity in self-supervised learning for surgical computer vision. In: MICCAI Workshop on Data Engineering in Medical Imaging. pp. 43–53. Springer (2024)
13. Jaspers, T.J., de Jong, R.L., Li, Y., Kusters, C.H., Bakker, F.H., van Jaarsveld, R.C., Kuiper, G.M., van Hillegersberg, R., Ruurda, J.P., Brinkman, W.M., et al.: Scaling up self-supervised learning for improved surgical foundation models. arXiv preprint arXiv:2501.09436 (2025)
14. Kassem, H., Alapatt, D., Mascagni, P., Karargyris, A., Padoy, N.: Federated cycling (fedcy): Semi-supervised federated learning of surgical phases. IEEE transactions on medical imaging **42**(7), 1920–1931 (2022)
15. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
16. 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. pp. 21074–21084 (2023)
17. Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al.: Surgical data science–from concepts toward clinical translation. Medical image analysis **76**, 102306 (2022)
18. Pérez, A., Rodríguez, S., Ayobi, N., Aparicio, N., Dessevres, E., Arbeláez, P.: Must: Multi-scale t ransformers for surgical phase recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 422–432. Springer (2024)
19. Shi, X., Jin, Y., Dou, Q., Heng, P.A.: Semi-supervised learning with progressive unlabeled data excavation for label-efficient surgical workflow recognition. Medical Image Analysis **73**, 102158 (2021)
20. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
21. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)

22. 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 transactions on medical imaging **36**(1), 86–97 (2016)
23. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
24. Xing, Z., Dai, Q., Hu, H., Chen, J., Wu, Z., Jiang, Y.G.: Svformer: Semi-supervised video transformer for action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18816–18826 (2023)
25. Xu, Y., Wei, F., Sun, X., Yang, C., Shen, Y., Dai, B., Zhou, B., Lin, S.: Cross-model pseudo-labeling for semi-supervised action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2959–2968 (2022)
26. Yang, S., Luo, L., Wang, Q., Chen, H.: Surgformer: Surgical transformer with hierarchical temporal attention for surgical phase recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 606–616. Springer (2024)