

MuST: Multi-Scale Transformers for Surgical Phase Recognition

Alejandra Pérez* , Santiago Rodríguez* , Nicolás Ayobi[◊], Nicolás Aparicio[◊], Eugénie Dessevres, and Pablo Arbeláez

Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Bogotá, Colombia
 {a.perezr20, s.rodriguezr2}@uniandes.edu.co

Abstract. Phase recognition in surgical videos is crucial for enhancing computer-aided surgical systems as it enables automated understanding of sequential procedural stages. Existing methods often rely on fixed temporal windows for video analysis to identify dynamic surgical phases. Thus, they struggle to simultaneously capture short-, mid-, and long-term information necessary to fully understand complex surgical procedures. To address these issues, we propose Multi-Scale Transformers for Surgical Phase Recognition (MuST), a novel Transformer-based approach that combines a Multi-Term Frame encoder with a Temporal Consistency Module to capture information across multiple temporal scales of a surgical video. Our Multi-Term Frame Encoder computes interdependencies across a hierarchy of temporal scales by sampling sequences at increasing strides around the frame of interest. Furthermore, we employ a long-term Transformer encoder over the frame embeddings to further enhance long-term reasoning. MuST achieves higher performance than previous state-of-the-art methods on three different public benchmarks.

Keywords: Surgical Workflow Analysis · Surgical Phase Recognition · Long-Term Transformers · Vision Transformers · Temporal Multi-Scale

1 Introduction

Surgical workflow analysis is critical for computer-assisted surgery, as it aims to understand the operational sequence of stages in surgical procedures [19,16,22]. Henceforth, equipping computer-aided systems with the ability to recognize these workflows would improve automated assistance to surgical teams, facilitate postoperative analysis, and contribute to medical personnel training [29]. An imperative step towards automatic surgical workflow analysis is identifying the succession of distinct phases throughout a surgical procedure [6]. Despite advancements in addressing this task, phase recognition remains challenging as it demands adaptable models capable of understanding the inherent variability in phase durations and the high semantic similarity between distinct phases present in surgical data [16]. Thus, these models need to incorporate long-term

* [◊] These authors contributed equally to this work

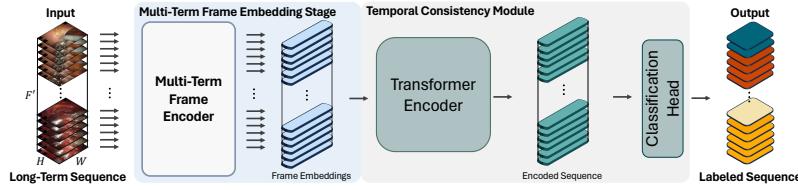


Fig. 1. MuST employs a Multi-Term Frame Encoder to generate rich embeddings containing short- and mid-term dependencies for a long-term sequence of F' frames. The Temporal Consistency Module introduces long-term analysis by processing relationships among frame embeddings to enable coherence in the predictions.

temporal context and ensure precise temporal consistency to achieve a reliable understanding of surgical progressions [18].

Pioneering works for surgical phase recognition used manually designed features to train classical machine learning classifiers [22,21]. However, Deep Learning methods gradually replaced them due to their superior capacity to represent temporal and spatial contexts. EndoNet [24] was the first approach that employed Convolutional Neural Networks (CNNs) to extract features that captured spatial dependencies but did not incorporate the temporal information from surgical videos. To address this limitation, PhaseNet [25], EndoLSTM [26], SV-RCNet [14], and OHFM [30] used long short-term memory (LSTM) networks to model temporal dependencies. Nevertheless, LSTM-based methods have limited long-term processing capacities due to their sequential nature and vanishing gradient issues. To mitigate this problem, TMRNet [15] proposed multi-scale non-local operations, but they needed to aggregate global information collaboratively. On the contrary, TeCNO [4] used Temporal Convolutional Networks (TCN) with dilated convolutions to achieve long-term temporal context, which limited its comprehension of short-term and fine-grained information.

Transformers [28] have proved to be remarkably powerful for sequential data and also for image and video analysis [2,9,31,8] due to their ability to model long-term dependencies and extract rich features through self-attention mechanisms. Multiple surgical workflow analysis models like OperA [5], SAHC [7], and Trans-SVNet [11] incorporated Transformer layers to TCNs in order to efficiently combine the spatial and temporal features. Nonetheless, their dependence on TCN modeling leads to a loss of finer-grained information, and using temporal-agnostic backbones limits frame embeddings to capture only spatial information. In contrast, TAPIR [27] and TAPIS [1] proved the utility of a fully Transformer-based model for multiple surgical workflow analysis tasks and improving frame representation by extracting rich spatio-temporal embeddings. Still, these models process short-term fixed-size windows and employ a per-frame prediction approach that limits its temporal consistency. Recently, SKiT [18] achieved long-term reasoning by using critical pooling to record relevant past information, but it also employs a backbone that only extracts spatial dependencies.

Several methods have studied the potential of multi-scale analysis to improve temporal comprehension in generic human activity recognition [10,23,12,20]. Initially, SlowFast [10] introduced the concept of sub-sampling video sequences at different rates to consider multiple temporal coverages. More recently, TemPr [23] extended this approach by using a set of Transformers on various sequences of increasing temporal coverages, enabling information aggregation from diverse scales. Still, these concepts need to be fully explored in surgical workflow analysis. Integrating them could be highly advantageous, as it introduces dynamic window designs for phase recognition that adapt to phases of varying lengths.

This work presents a novel approach to augment flexibility over surgical phase variability by processing multiple temporal scales. For this purpose, we introduce Multi-Scale Transformers for Surgical Phase Recognition (MuST), a fully Transformer-based model composed of two processing stages. The first stage corresponds to a Multi-Term Frame Encoder (MTFE) that captures information from short- and mid-term temporal scales into multi-term frame embeddings. Within this stage, we create a temporal pyramid comprising multiple sequences sampled at progressively increasing rates [23]. Inspired by multi-scale image processing Transformers [3], we introduce a Multi-Temporal Attention Module that correlates information from each sequence and combines patterns across short- and medium-term periods, offering a flexible understanding of the information wrapping each frame. The second stage is a Temporal Consistency Module (TCM) that employs a lightweight and long-term Transformer encoder to process extensive sequences of multi-term frame embeddings. This approach enables our TCM to encode long-term dependencies within the surgical procedure and predict temporally coherent surgical phase segments. Hence, the MTFE and the TCM complement each other to achieve holistic temporal modeling of complex surgical procedures.

To sum up, our main contributions are: (1) We introduce a multi-sequence pyramid and a Multi-Temporal Attention Module for surgical phase recognition, allowing the combined analysis of temporal windows with a hierarchical temporal coverage over a surgical video, and (2) we propose a temporal consistency module that leverages Transformers’ attention over the extracted frame embeddings to improve the consistency along wide-range temporal segments.

We demonstrate MuST’s superiority over previous state-of-the-art models by extensively evaluating on three public benchmarks. To ensure reproducibility, all our source code and pretrained models are available at <https://github.com/BCV-Uniandes/MuST>.

2 Method

We present **Multi-Scale** Transformers for Surgical Phase Recognition (MuST), a two-stage Transformer-based architecture designed to enhance the modeling of short-, mid-, and long-term information within surgical phases. Our method employs a frame encoder that leverages multi-scale surgical context across different temporal dimensions. The frame encoder considers diverse time spans

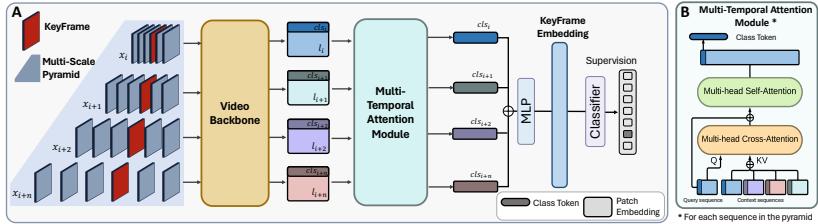


Fig. 2. MuST’s Multi-Term Frame Encoder utilizes MViT to capture embeddings of a temporal sequence sampled at multiple rates (x_i) with the keyframe placed at the middle position for offline setups and at the last position for online setups. Each sequence embedding (l_i), along with their corresponding class token (cls_i), goes through the Multi-Temporal Attention Module with the rest of the sequence embeddings of the pyramid as temporal context. A final rich multi-term embedding of the keyframe is generated by concatenating class tokens from each scale and processing them through a multi-layer perceptron (MLP).

around a specific frame of interest, which we call a keyframe. The keyframe serves as the specific frame that we encode. We construct temporal windows around this keyframe to provide the necessary temporal context for accurate phase classification. Our encoder generates rich embeddings that capture short- and mid-term dependencies. To further enhance long-term understanding, we employ a Temporal Consistency Module (TCM) that establishes relationships among frame embeddings within an extensive temporal window, ensuring coherent phase recognition within an extensive temporal window.

As illustrated in Fig. 2, MuST begins by constructing a pyramid of sequences around each keyframe, capturing a hierarchy of temporal scales with increasing sample rates [23]. Our method leverages a Video Backbone [9] to obtain rich feature representations of the temporal context around a surgical keyframe provided by each sampled sequence. Then, our Multi-Temporal Attention Module captures interdependencies between the different temporal scales and retains the class token from each sequence. Furthermore, we concatenate the class tokens through a Multi-Layer Perceptron (MLP) to generate rich multi-term embeddings. Finally, our TCM self-attends the generated frame embeddings from the MLP and enhances coherence in predictions across an extensive temporal window, as shown in Fig. 1.

Temporal Multi-Scale Pyramid. We sample sub-sequences from a surgical video at increasing rates to construct a temporal multi-scale pyramid. In offline setups, each sequence places the keyframe in the middle position, whereas in online setups, it is placed at the last position. Thus, this pyramid offers finer-grained temporal detail with lower temporal coverage at the first levels, while the last levels provide broader but sparse temporal context. This hierarchical structure enables simultaneous analysis at multiple temporal resolutions and captures information efficiently across various time scales.

Given a total number of N sampling scales, we define the set of sampled subsequences as $X = \{x_i\}_i^N$ where each sequence $x_i \in \mathbb{R}^{T \times H \times W \times 3}$ has T RGB frames of $H \times W$ pixels and the temporal sampling stride of a sequence x_i increases as i approaches N . For each input sub-sequence, we use a shared video backbone $\mathcal{M}(\bullet)$ to compute a sequence of spatio-temporal embeddings $l_i \in \mathbb{R}^{T' \times D}$ with T' embeddings of length D . We define the set of output embeddings from the video backbone as $L = \{l_i\}_i^N$ where $l_i = \mathcal{M}(x_i)$.

Video Backbone. We adopt MViT [9] as our video backbone. MViT processes a fixed temporal window throughout multiple stages containing several Transformer blocks. At the beginning of each stage, a Multi-Head Pooling Attention (MHPA) mechanism reduces the space-time resolution while augmenting the feature dimension. Thus, MViT produces an implicit multi-scale feature pyramid where the initial stages capture detailed spatial resolutions while the final stages encode a shorter sequence of complex spatio-temporal embeddings. Additionally, MViT concatenates a learnable class token to capture a single embedding of the entire sequence. For further details, we refer the reader to the original MViT [9] paper. MuST retains the class token and the patch embeddings from the video backbone for each pyramid level.

Multi-Temporal Attention Module. We introduce a novel Multi-Temporal Attention Module to compute relationships among the spatio-temporal representations extracted from each pyramid scale [3], as detailed in Fig. 2B. This module is composed of two units: a Multi-Temporal Cross-Attention (MTCA) unit to explore sequence inter-dependencies, denoted as $MTCA(\bullet)$, and a residual Multi-Temporal Self-Attention (SA) unit to identify relationships among all computed embeddings, symbolized by $SA(\bullet)$. The $MTCA$ module computes cross-attention between each embedding sequence with the concatenation of all sequences. For the i -th set of embeddings generated from the i -th sub-sequence, we compute $MTCA$'s attention as follows,

$$l' = \text{concat}(L), \quad Q_i = W_i^Q l_i, \quad K_i = W_i^K l'_i, \quad V_i = W_i^V l'_i \quad (1)$$

$$\mathbf{MTCA}(l_i, l'_i) = \text{softmax} \left(\frac{Q_i \cdot K_i^T}{\sqrt{d_{k_i}}} \right) V_i \quad (2)$$

Where $\text{concat}(L)$ represents the concatenation operation across the sequence axis and W_i^Q , W_i^K , and W_i^V represent the learnable weights of the queries (Q), keys (K), and values (V) for the i -th embedding, respectively. We designate the set of output embeddings from the $MTCA$ by $C = \{c_i\}_i^N$ where $c_i = \text{MTCA}(l_i, l'_i)$ and $c_i \in \mathbb{R}^{T' \times D}$. Further on, for each c_i we calculate self-attention in the SA module as follows:

$$c'_i = \text{concat}(l_i, C), \quad Q_i = W_i^Q c'_i, \quad K_i = W_i^K c'_i, \quad V_i = W_i^K c'_i \quad (3)$$

$$\mathbf{SA}(c'_i) = \text{softmax} \left(\frac{Q_i \cdot K_i^T}{\sqrt{d_{k_i}}} \right) V_i \quad (4)$$

Where $\text{concat}(l_i, C)$ represents the concatenation across the sequence dimension of all the embeddings in C with the l_i sequence produced by the video backbone as a residual connection, thus $c'_i \in \mathbb{R}^{(N+1)T' \times D}$. Finally, we concatenate all class tokens into a single embedding $p' \in \mathbb{R}^{ND}$ which is linearly transformed by an MLP into the final multi-term frame embedding denoted as $p \in \mathbb{R}^{ND}$ where $p = \text{MLP}(p')$. We train our shared video backbone and our Multi-Temporal Attention Module by adding a linear classifier that projects each p into a class probability distribution corresponding to the phase class of the middle frame in the multi-sequence pyramid.

Temporal Consistency Module (TCM). We adopt Transformers’ self-attention to capture relationships between multiple frames and enhance their understanding of long-term dependencies. Given a video $V \in \mathbb{R}^{F \times H \times W \times 3}$ with F frames, we construct overlapping windows $v = \{p_j\}_j^{F'}$ of F' frames ($F' < F$) where p_j corresponds to the multi-term frame embedding generated for the j -th frame in the window. Thus, the total number of windows in v will be equal to $1 + \frac{F-F'}{F'-Overlap}$. We perform self-attention as follows:

$$\mathbf{v}' = \text{concat}(v) + \text{PE}, \quad \mathbf{Q} = W^Q v', \quad \mathbf{K} = W^K v', \quad \mathbf{V} = W^V v' \quad (5)$$

$$\text{TCM}(v) = \text{softmax} \left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V} \quad (6)$$

Where $\text{concat}(v)$ is the concatenation of all embeddings in v into a sequence of F' embeddings, and PE corresponds to a cosine positional embedding. Finally, we linearly project each output embedding into a phase class distribution. For offline inference, we average all the probability distributions obtained for each frame throughout all sampled overlapping windows.

3 Experiments and Results

3.1 Experimental Setup

Datasets. We evaluate the performance of MuST in three surgical datasets: HeiChole [29], GraSP [1], an extended version of PSI-AVA [27], and MISAW [13]. These datasets are standard benchmarks in surgical workflow analysis from different domains. We adhere to each dataset’s public benchmarks and metrics to ensure fair comparison with other models. Thus, we evaluate using the F1-score in HeiChole in an online setup and mean Average Precision (mAP) in GraSP and MISAW offline. Since HeiChole’s test data is not public, we used the training and validation sets proposed by [29] for comparison with other methods. For GraSP, we perform ablation studies using its cross-validation set and report final results using its test set.

Implementation Details. We initialize the weights from an MViT-B model pretrained on phase recognition tasks for 20 epochs. Using these pretrained weights, we train the MuST Multi-Term Encoder for an additional five epochs,

utilizing three NVIDIA Quadro RTX 8000 GPUs with a batch size of 18. For the offline setup, we construct the Temporal Multi-Scale Pyramid using four sequences of 16 frames each, with sampling rates of 1, 4, 8, and 12 seconds. We use 24 frames with the same sampling rates for the online setup. Specifically for HeiChole, we adapt the temporal window to include only past frames to ensure a fair comparison with other online methods. We train the Temporal Consistency Module (TCM) for 20 epochs on a single GPU with a batch size of 256. The temporal window covers 10% of the mean video duration for offline datasets and 5% for online datasets, with a 90% overlap for all datasets. The training for all modules in MuST utilizes cross-entropy loss, optimized using AdamW with a cosine learning rate scheduler, a weight decay of 1×10^{-4} , and an initial learning rate of 1×10^{-4} . Our model comprises 68 million parameters.

3.2 Experimental Validation

We show the overall performance in Table 1 and our qualitative comparisons in Fig. 3. We compare our model with open-source, state-of-the-art methods. We train and optimize these methods on our selected benchmarks to ensure fair and accurate comparisons. MuST outperforms previous state-of-the-art methods across all datasets, achieving excellent results in both online and offline setups. These results highlight the effectiveness of multi-scale temporal reasoning in ensuring consistent predictions across various temporal intervals.

Table 1. Comparative Results of MuST in different benchmarks. The best results are shown in bold. All methods were adapted to conduct offline or online inference according to the dataset.

a) Grasp		b) MISAW		c) HeiChole	
Model	Phases mAP	Model	Phases mAP	Model	F1-score
TAPIS	76.07	TAPIS	97.14	TAPIS	73.41
TeCNO	77.10	TeCNO	95.58	TeCNO	69.35
Trans-SVNet	76.54	Trans-SVNet	90.38	Trans-SVNet	71.85
MuST (ours)	79.14	MuST (ours)	98.08	MuST (ours)	77.25

Our method consistently overperforms TAPIS, which employs an MViT backbone with a fixed temporal window. Compared to TeCNO, MuST improves phase recognition metrics in short- and medium-duration time intervals. We present per-class performance and phase mean durations across all datasets in Table 1 of the Supplementary Material. These results prove the advantages of incorporating spatio-temporal reasoning of sequences facilitated by MuST’s backbone instead of per-frame embeddings coming from a CNN classifier. Along with the Multi-Temporal Attention Module and the TCM, MuST distinguishes phase time intervals and durations more accurately.

Finally, compared to Trans-SVNet, which relies on short-term fixed-size windows and per-frame prediction, MuST effectively captures rapid phase changes,

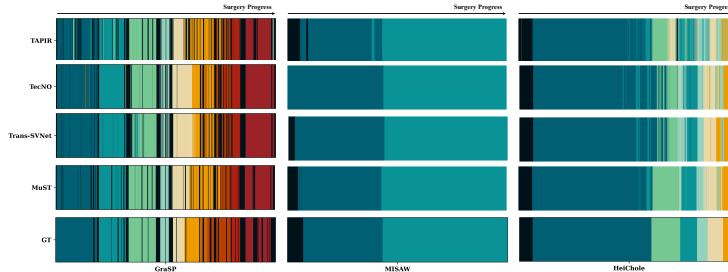


Fig. 3. Qualitative results. Qualitative comparisons with state-of-the-art methods on one representative video from each dataset.

particularly evident in the GraSP dataset (Fig. 3), where phases exhibit highly variable time durations. This capability arises from the multi-term reasoning from the MTFE. This encoder learns relationships within each sub-sequence in the pyramid, producing rich multi-term frame embeddings that are subsequently refined by the TCM for segment predictions that relate a wider temporal context within the sequence. Furthermore, we conducted an additional validation of MuST on the Cholec80 dataset [24] and obtained an F1-score of 85.57%, precision of 84.6%, and recall of 86.6%, achieving a highly competitive performance, comparable to current state-of-the-art methods on this benchmark [18,17].

3.3 Ablation Experiments

We conduct ablations on the cross-validation set of the GraSP benchmark, as detailed in Table 2 and displayed in Fig. 3 of the Supplementary Material. Our experiments demonstrate that using multiple sequences sampled at the same rate as input leads to a performance increase, likely due to higher redundancy that allows a wider understanding of the patterns present in the data. Similarly, we observe a significant performance boost when incorporating the Temporal Multi-Scale Pyramid, which constructs windows at varying sample rates. This approach aggregates information from multiple scales, resulting in richer contextual information. Adding cross-attention and self-attention mechanisms further improves the model’s performance as they establish cross-scale reasoning that enables interactions between sequences in the pyramid. This is evident in the cross-attention maps in Fig. 2 of the Supplementary Material. Lastly, the TCM significantly enhances the model’s prediction coherence by incorporating long-term dependencies between frame embeddings through its transformer encoder architecture. This improvement is illustrated in Fig. 3 of the Supplementary Material.

4 Conclusions

In this work, we introduce a novel method for surgical phase recognition that effectively addresses the challenges associated with varying phase durations in

Table 2. Ablation experiments results for MuST in the GraSP benchmark.

The Multi-Sequence column indicates if the model processes multiple input windows, and the Temporal Multi-Scale Pyramid column denotes the use of multiple temporal scales. The Multi-Temporal Attention Module column indicates the use of attention mechanisms across multiple temporal scales. TCM represents the inclusion of the Temporal Consistency Module. A checkmark (✓) indicates the presence of an attribute, while a blank space denotes its absence. The best result is highlighted in bold.

Multi-Sequence	Temporal Multi-Scale Pyramid	Multi-Temporal Attention Module	TCM	mAP
✓				72.57 ± 0.09
✓	✓			73.38 ± 2.30
✓	✓	✓		75.87 ± 2.19
✓	✓	✓	✓	76.13 ± 1.74
				77.34 ± 1.02

surgical data. By leveraging a temporal pyramid and a cross-attention module, we enrich the temporal context and facilitate multi-scale learning, capturing short- and mid-term dependencies. Additionally, we present a Temporal Consistency Module to enhance long-term reasoning and strengthen the model’s overall performance. We demonstrate that MuST outperforms previous state-of-the-art methods across different surgical datasets and establishes a flexible framework for future research in complex temporal modeling in surgical settings.

Acknowledgments. Alejandra Pérez, Santiago Rodriguez, and Nicolás Ayobi acknowledge the support of the 2022 and 2023 UniAndes-DeepMind Scholarships.

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

References

1. Ayobi, N., Rodríguez, S., Pérez, A., et al.: Pixel-wise recognition for holistic surgical scene understanding (2024)
2. Carion, N., Massa, F., Synnaeve, G., et al.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 357–366 (2021)
4. Czempiel, T., Paschali, M., Keicher, M., et al.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020. pp. 343–352. Springer (2020)
5. Czempiel, T., Paschali, M., Ostler, D., et al.: Opera: Attention-regularized transformers for surgical phase recognition. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021. pp. 604–614. Springer (2021)
6. Demir, K.C., et al.: Deep learning in surgical workflow analysis: A review of phase and step recognition. IEEE Journal of Biomedical and Health Informatics **27**(11), 5405–5417 (2023). <https://doi.org/10.1109/JBHI.2023.3311628>

7. Ding, X., Li, X.: Exploring segment-level semantics for online phase recognition from surgical videos. *IEEE Transactions on Medical Imaging* **41**(11), 3309–3319 (2022)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Fan, H., et al.: Multiscale vision transformers. In: international conference on computer vision. pp. 6824–6835 (2021)
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6201–6210 (2019). <https://doi.org/10.1109/ICCV.2019.00630>
11. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021. pp. 593–603. Springer (2021)
12. Girdhar, R., Grauman, K.: Anticipative Video Transformer. In: ICCV (2021)
13. Huaulmé, A., et al.: Micro-surgical anastomose workflow recognition challenge report. *Computer Methods and Programs in Biomedicine* **212**, 106452 (2021)
14. 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 (2017)
15. Jin, Y., Long, Y., Chen, C., et al.: Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Medical Imaging* **40**(7), 1911–1923 (2021)
16. Kirtac, K., Aydin, N., Lavanchy, J.L., et al.: Surgical phase recognition: From public datasets to real-world data. *Applied Sciences* **12**(17), 8746 (2022)
17. 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. arXiv preprint arXiv:2305.08989 (2023)
18. Liu, Y., Huo, J., Peng, J., et al.: 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)
19. Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., et al.: Surgical data science – from concepts toward clinical translation. *Medical Image Analysis* **76**, 102306 (2022). <https://doi.org/https://doi.org/10.1016/j.media.2021.102306>, <https://www.sciencedirect.com/science/article/pii/S1361841521003510>
20. Meng, Y., et al.: Ar-net: Adaptive frame resolution for efficient action recognition. In: Computer Vision–ECCV 2020. pp. 86–104. Springer (2020)
21. Padoy, N.: Machine and deep learning for workflow recognition during surgery. *Minimally Invasive Therapy & Allied Technologies* **28**(2), 82–90 (2019)
22. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. *Medical Image Analysis* **16**(3), 632–641 (2012). <https://doi.org/https://doi.org/10.1016/j.media.2010.10.001>, <https://www.sciencedirect.com/science/article/pii/S1361841510001131>, computer Assisted Interventions
23. Stergiou, A., Damen, D.: The wisdom of crowds: Temporal progressive attention for early action prediction. In: Conference on Computer Vision and Pattern Recognition. pp. 14709–14719 (2023)
24. Twinanda, A.P., et al.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging* **36**(1), 86–97 (2016)

25. Twinanda, A.P., et al.: Single-and multi-task architectures for surgical workflow challenge at m2cai 2016. arXiv preprint arXiv:1610.08844 (2016)
26. Twinanda, A.P.: Vision-based approaches for surgical activity recognition using laparoscopic and RGBD videos. Ph.D. thesis, Strasbourg (2017)
27. Valderrama, N., Ruiz Puentes, P., Hernández, I., Ayobi, N., Verlyck, M., et al.: Towards holistic surgical scene understanding. In: International conference on medical image computing and computer-assisted intervention. pp. 442–452. Springer (2022)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
29. Wagner, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Medical Image Analysis **86**, 102770 (2023)
30. Yi, F., Jiang, T.: Hard frame detection and online mapping for surgical phase recognition. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019. pp. 449–457. Springer (2019)
31. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.a.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)

MuST: Supplementary Material

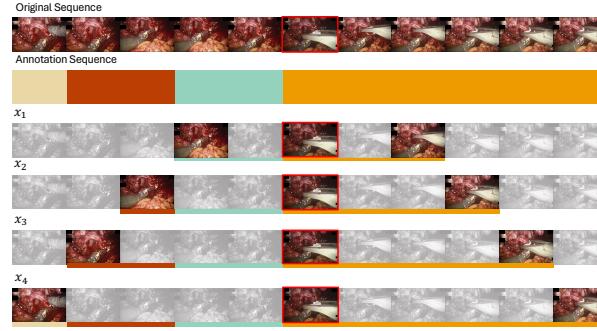


Fig. 1. Temporal Multi-Scale Pyramid Example. Illustration of a temporal multi-sequence pyramid centered on the red keyframe. Sampled frames at each level are highlighted. The top levels of the pyramid capture short-term temporal information, while the lower levels comprehend a broader temporal context.

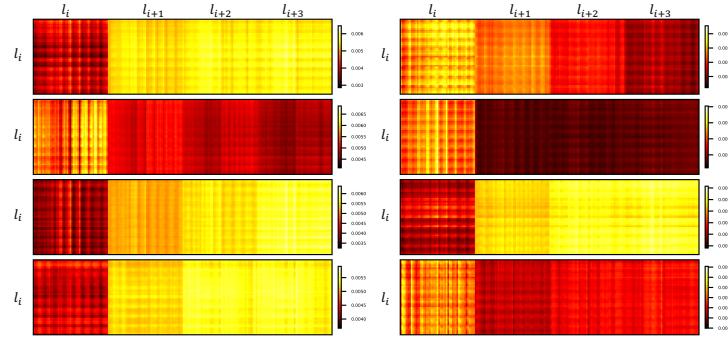


Fig. 2. Cross-attention maps. Attention weights from the 8 attention heads of one of the layers of the cross-attention described in Section 2 (Method) for the l_i sequence of the set L . These attention maps illustrate that each head specializes in capturing different segments of the L set, featuring inter-dependencies across temporal scales.

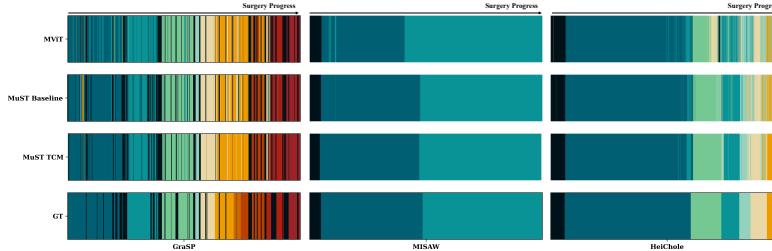


Fig. 3. MuST visualization performance in every benchmark. MuST baseline refers to our model without the Temporal Consistency Module (TCM), whereas MuST TCM represents the complete model. Overall visualizations demonstrate that MuST achieves consistent recognition of phases, including reclassification of some segments when using the Temporal Consistency Module.

Table 1. Class comparison of MuST with state-of-the-art methods. MuST baseline refers to our model without the Temporal Consistency Module (TCM), whereas MuST TCM represents the complete model. Consistent results across classes are evident across diverse temporalities within the databases. These results underscore MuST’s proficiency in identifying multiple duration phases, attributed to its multi-temporal reasoning. Leveraging information from distinct temporal contexts enhances MuST’s predictive capabilities for phases of different durations. The best results are highlighted in bold.

GraSP						
Phase	Duration	TAPIS	TeCNO	Trans-SVNet	MuST Baseline (Ours)	MuST TCM (Ours)
0	50.60 ±107.06	85.5	77.9	74.7	86.8	87.1
1	285.28±318.04	86.0	95.8	94.2	95.6	97.7
2	221.55±256.20	76.0	87.7	84.5	91.6	93.0
3	143.33±159.81	83.0	83.2	85.3	91.9	94.2
4	84.73±55.18	83.8	67.1	76.1	90.8	89.5
5	149.16±149.92	85.0	78.5	82.3	89.5	87.8
6	222.21±264.64	73.5	81.6	79.4	79.4	82.1
7	108.00±56.66	8.6	68.1	61.1	16.6	22.6
8	74.89±76.06	71.6	70.1	66.1	71.8	76.2
9	107.00±77.22	59.3	70.2	67.9	63.8	60.4
10	70.48±80.06	76.3	67.8	70.2	78.9	79.8

MISAW						
Phase	Duration	TAPIS	TeCNO	Trans-SVNet	MuST (Ours)	MuST TCM (Ours)
0	114.70±62.62	92.1	87.8	75.7	94.2	95.0
1	2318.22±931.10	97.9	99.4	96.6	98.9	99.5
2	3651.41±2759.67	98.9	99.6	98.8	99.4	99.8

HeiChole						
Phase	Duration	TAPIS	TeCNO	Trans-SVNet	MuST (Ours)	MuST TCM (Ours)
0	170.38±63.18	86.4	89.3	92.3	96.2	97.6
1	618.67±461.98	83.0	81.9	84.2	92.3	91.4
2	127.85±61.96	46.4	37.8	40.3	61.6	58.0
3	384.68±222.82	73.8	79.2	78.1	76.6	77.6
4	86.33±37.20	72.1	64.1	75.4	67.2	68.5
5	279.48±208.76	58.5	65.6	67.7	70.0	73.2
6	70.12±63.01	64.7	69.3	64.9	67.9	73.0