Trans-SVNet: Hybrid Embedding Aggregation Transformer for Surgical Workflow Analysis

Yueming Jin[1], Yonghao Long[2], Xiaojie Gao[2], Danail Stoyanov[1], Qi Dou[2*] and Pheng-Ann Heng[2]

[1]Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), Department of Computer Science, University College London, UK.

[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, HK, China.

*Corresponding author. E-mail: qidou@cuhk.edu.hk.

Abstract

Purpose: Real-time surgical workflow analysis has been a key component for computer assisted intervention system to improve cognitive assistance. Most existing methods solely rely on conventional temporal models and encode features with a successive spatial-temporal arrangement. Supportive benefits of intermediate features are partially lost from both visual and temporal aspects. In this paper, we rethink feature encoding to attend and preserve the critical information for accurate workflow recognition and anticipation. **Methods:** We introduce Transformer in surgical workflow analysis, to reconsider complementary effects of spatial and temporal representations. We propose a hybrid embedding aggregation Transformer, named Trans-SVNet, to effectively interact the designed spatial and temporal embeddings, by employing spatial embedding to query temporal embedding sequence. We jointly optimized by loss objectives from both analysis tasks to leverage their high correlation. **Results:** We extensively evaluate our method on three large surgical video datasets. Our method consistently outperforms the state-of-the-arts across three datasets on workflow recognition task. Jointly learning with anticipation, recognition results can gain a large improvement. Our approach also shows its effectiveness on anticipation with promising performance achieved. Our model achieves a real-time inference speed of 0.0134 second per frame. **Conclusion:** Experimental results demonstrate the efficacy of our hybrid embeddings integration by rediscovering the crucial cues from complementary spatial-temporal embeddings. The better performance by multi-task learning indicate that anticipation task brings the additional knowledge to recognition task. Promising effectiveness and efficiency of our method also show its promising potential to be used in operating room.

Keywords: Surgical vision, workflow recognition, workflow anticipation, Transformer, spatial-temporal feature modeling

1 Introduction

Intelligent computer assisted intervention (CAI) has greatly enhanced the safety and quality of patient care in modern operating theatres [1]. Surgical workflow analysis (SWA), such as surgical workflow recognition and anticipation, is fundamentally essential for the CAI system [2, 3]. Workflow recognition aims at identifying the current surgical phase, step, or activity, with different granularities, while workflow anticipation focuses on predicting the future event, where recent works formulate it to estimate the remaining surgery time until the occurrence of a specific event [4, 5]. These two SWA tasks enable CAI systems to monitor surgical process [3], support the context-aware decision making [2], facilitate communication in operating room [6], and reduce deviations and anomalies by providing the early warning [1]. They can also assist the preparation before the next surgical phase, which further helps development of intelligent robotic system and surgery automation, e.g., automatically triggering the instrument preparation for incoming intervention [4]. However, pure vision-based workflow analysis is highly challenging. Surgical scene is complicated, leading to similar inter-phase appearance and high intra-phase variance. Surgical phase transitions are ambiguous in different procedures, due to various operative skills and patient cohorts. The motion blur, lighting condition changes, and noise also increase understanding difficulty.

Temporal information modeling is key ingredient for accurate analysis. To solve above challenges, a set of approaches have been developed to explore temporal clues. Early methods utilize linear statistical models, such as variants of hidden Markov models [7], and Dynamic Time Warping (DTW) [8]. Deep networks have achieved remarkable progresses in workflow analysis. One stream of works seamlessly integrate convolutional neural network (CNN) and long short-term memory (LSTM) network to model high-level spatio-temporal representations [4, 9–11]. 3D CNN has also been proposed to explore spatio-temporal cues with high-dimensional kernels [12]. However, these methods only collect information from short-term temporal span. Recently, several studies employ Temporal Convolutional Networks (TCN) [13] with dilated kernels

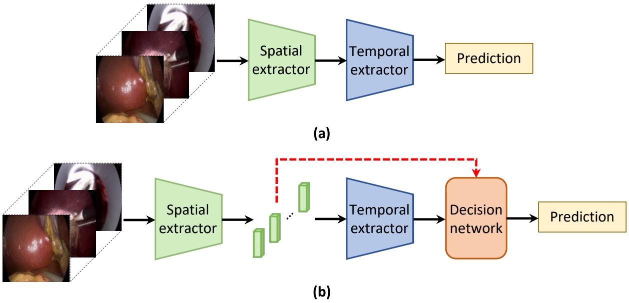


Fig. 1 (a) Previous methods solely rely on conventional temporal models and extract spatio-temporal features successively; (b) We propose to reuse extracted spatial features with temporal features to improve results.

to leverage longer-range dynamics, achieving state-of-the-arts on both recognition [14] and anticipation [5] tasks. However, some temporal cues may be inevitably lost due to the dilation. Apart from suboptimal temporal information capturing, we identify another limitation in current literature. As shown in Fig. 1 (a), existing methods extract spatial and temporal features successively, leading to losses of critical spatial attributes in final feature representations for prediction.

Transformer, with self-attention mechanism, enables a different perspective for representation encoding. It is originally introduced for natural language processing [15], and has revolutionized various vision tasks recently [16]. It allows relating each entry inside a sequence at different positions, named key embeddings, to a query embedding. Therefore, it can preserve long-term clues hampered in LSTM/3D CNN-based methods, and also sufficiently encode temporal cues lost in TCN-based methods. Moreover, the separate input of key and query in Transformer provides opportunity to simultaneously consider different kinds of features, whose efficacy has been demonstrated in various tasks [17, 18], such as point cloud registration by fusing multiple views [18]. It implies its potential to promote the synergy of spatial and temporal features for accurate workflow analysis.

Additionally, given high correlation among different tasks in surgical video understanding, multi-task learning has shown great effectiveness, such as joint phase and instrument recognition [19], gesture recognition and skill assessment [20]. However, simultaneously recognizing and anticipating workflow is rarely explored. One related work [21] utilizes DTW to jointly predict current phase and next one. Another [4] designs probabilistic CNN-LSTM for instrument anticipation, using tool recognition for additional regularization.

In this paper, we propose a novel hybrid embedding aggregation **Trans**former, for joint workflow recognition and anticipation from **S**urgical **V**ideos, named Trans-SVNet. Specifically, we introduce Transformer for surgical workflow analysis, to leverage its superior capability of sequential information modeling. We further reconsider the spatial features as one of the hybrid embeddings to supply missing appearance details during temporal feature extracting (Fig. 1 (b)). It is achieved by regarding spatial embeddings

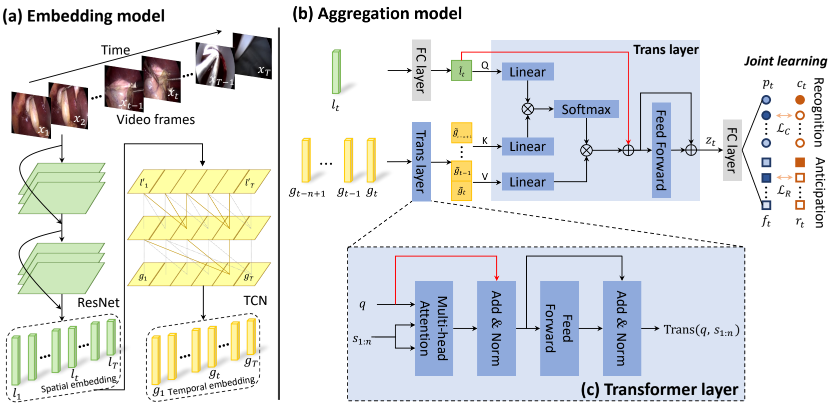


Fig. 2 Overview of our proposed Trans-SVNet. (a) Extracted spatial embeddings enable generation of temporal embeddings. (b) Our Transformer-based hybrid aggregation model is jointly optimized by two objectives. (c) Transformer layer in detail.

as the query, to attend the supporting information from temporal embedding sequences as the key. Additionally, the proposed model is jointly learned with two objectives to leverage the high relatedness of two tasks. We extensively evaluate our Trans-SVNet on three large surgical video datasets, including laparoscopic and cataract procedures. Our method consistently outperforms the state-of-the-arts across all three datasets for workflow recognition. Meanwhile, our Transformer-based model attains competitive and promising results for workflow anticipation. It is interesting to find that simple multi-task learning can already promote recognition results, by revealing the complementary benefits from anticipation without extra annotation efforts. Our model achieves a high inference speed, showing extraordinary potential for real-time applications.

A preliminary version of this work was presented in MICCAI 2021 [22]. In this paper, we have substantially revised and extended the original version. The main modifications include providing a more comprehensive literature review; extending method from tackling single task to joint learning of two tasks; elaborating methods in details; adding one more dataset of different surgery for evaluation; discussing potential in clinical usage and pointing out some future remedies. Code will be publicly available.

2 Methods

Overview of our proposed Trans-SVNet is presented in Fig. 2. In this section, we first show embedding extraction models to represent video frames by spatial embedding and temporal embedding. Then we describe the architecture of utilized Transformer layer for relational feature aggregation, and our proposed aggregation model to explore the synergy of hybrid embeddings for accurate workflow analysis. Joint learning towards two objectives are finally developed for simultaneous workflow recognition and anticipation.

2.1 Spatial and Temporal Embedding Extraction

Video frames in surgical procedure inherently contain two types of essential information, i.e., spatial information within each single frame, and temporal nature after considering other frames. Therefore, we extract two types of embeddings to describe the spatial and temporal information respectively. We denote $x_t \in \mathbb{R}^{H\times W\times C}$ as the t-th frame of a surgical video with T frames in total, and $y_t \in \mathbb{R}^N$ is corresponding ground truth with N dimensions considering both recognition and anticipation tasks. Regarding the spatial information, x_t is forwarded to a standard CNN model to extract discriminative spatial embedding (a 50-layer deep residual network ResNet50 [23] in our work). Followed by a frame-wise classifier, the spatial embedding extraction model is trained towards ground truths of our two tasks, i.e., workflow recognition and anticipation. The output vector of the average pooling layer in ResNet50 are exploited as the spatial embedding $l_t \in \mathbb{R}^{2048}$.

We further extract temporal embeddings based on spatial embeddings generated by ResNet50, to save memory and time. Concretely, we first reduce the dimension of l_t to $l'_t \in \mathbb{R}^{32}$ by using a 1D convolutional layer with kernel size 1. Then we input the entire video sample, i.e., spatial embeddings of all frames in a video $l'_{1:T}$, to a series of temporal convolutional layers, to yield temporal embeddings $g_{1:T}$. The longer-term multi-scale dependencies can be captured in each temporal embedding g_t. Meanwhile, process sticks to an online mode without accessing future information for practical usability. Here, we leverage TeCNO [14], a two-stage online TCN for generation. The whole model is trained towards ground truths with the parameters of ResNet50 keeping freezing. Outputs of the last stage in TeCNO are exploited as temporal embedding $g_t \in \mathbb{R}^N$, which can show a strong spatial-temporal representation by encoding long-range compositional motion information.

2.2 Transformer Layer for Relational Aggregation

Different from LSTM and TCN when aggregating temporal information, Transformer leverages the self-attention mechanism to augment each frame. It relates the information of all frames within a sequence (key samples), to the current one (query sample) in parallel, resulting in an information-preserved yet computation-efficient aggregation procedure. We introduce our utilized Transformer layer in this section. See Fig. 2 (c), inputting an embedding of a single frame q as the query, and embeddings of a sequence $s_{1:n} = [s_1, \ldots, s_{n-1}, s_n]$ as the key and value, the multi-head attention layer first calculates the relation between query and key, and value information is aggregated to the query based on the relation:

$$\mathrm{Attn}(q, s_{1:n}) = \mathrm{softmax}(\frac{W_q q (W_k s_{1:n})^{\mathrm{T}}}{\sqrt{d_k}}) W_v s_{1:n}, \quad (1)$$

where $\{W_q, W_k, W_v\}$ are linear mapping matrices and d_k is the dimension of q after linear transformation. Since each attention head owns it distinct learnable parameters, concentrating on its respective interests, the outputs of all heads are concatenated to convey the joint knowledge. It then connects with q in a residual way, followed by layer normalization. Next, features are fed into a feed-forward layer, also followed by residual adding and layer normalization. We denote one Transformer layer function as $\text{Trans}(q, s_{1:n})$, whose final output is the augmented representation of q after relating information from $s_{1:n}$.

2.3 Hybrid Embedding Aggregation

Most existing works using Transformer exploit different data samples as query and key, but with the same level representations [16]. Generally, they utilize the highest-level representations generated by feature extraction model, and for our task, is the strong spatial-temporal feature g_t. However, we argue that the temporal embedding g_t shall inevitably lose some inherent spatial context of each individual frame, after TCN refinement. Instead of completely relying on g_t to represent both spatial context and temporal dependency, we recall spatial embedding l_t to support pure spatial content. We then design a hybrid embedding aggregation model to leverage both-level representations by treating them as query and key, which allows the rediscovery of missing yet crucial spatial details.

Specifically, our aggregation model consists of two Transformer layers, aiming to generate the superior representation z_t of frame x_t by fusing the precomputed hybrid video embeddings. Before integrating the two embeddings, both of them first perform the internal aggregation respectively. For the spatial embedding, the dimension reduction is conducted to generate $\tilde{l}_t \in \mathbb{R}^N$ by

$$\tilde{l}_t = \tanh(W_l l_t), \tag{2}$$

where $W_l \in \mathbb{R}^{N \times 2048}$ is a parameter matrix. For the temporal embedding, we utilize an n-length sequence $g_{t-n+1:t}$ when analyzing the current t-th frame. The temporal embedding sequence $g_{t-n+1:t}$ is first forwarded to one Transformer layer to conduct the self-aggregation, obtaining an intermediate sequence $\tilde{g}_{t-n+1:t} \in \mathbb{R}^{n \times N}$. In other words, each entry in $[g_{t-n+1}, \ldots, g_{t-1}, g_t]$ attends all entries of the sequence as:

$$\tilde{g}_i = \text{Trans}(g_i, g_{t-n+1:t}), \quad i = t-n+1, \ldots, t. \tag{3}$$

Given self-aggregated embeddings $\tilde{l}$ and $\tilde{g}$, we employ the other Transformer layer with same architecture to interact the message of two embeddings. See Fig. 2 (b), $\tilde{l}_t$ queries the pivotal information from $\tilde{g}_{t-n+1:t}$ as key and value. Note that in Transformer layer, query is fused after multi-head attention (red arrow), the purified spatial embedding therefore can be fused through residual addition to advance the visual knowledge proportion. The final output of our aggregation model $z_t \in \mathbb{R}^N$ contains the hybrid message associating both space and time dimensions via one Transformer layer:

$$z_t = \text{Trans}(\tilde{l}_t, \tilde{g}_{t-n+1:t}). \tag{4}$$

2.4 Overall Loss Function for Joint Analysis

Considering the synergistic knowledge of space and time cues is crucial for both recognition and anticipation tasks. To leverage relationship, we jointly train the model towards loss objectives of both tasks. The strong representation z_t is fed into one fully-connected layer for feature transformation, generating phase classification probability p_t, and remaining time value f_t.

For workflow recognition, ground truth $c_t \in \mathbb{R}^{N_c}$ is one-hot vector with phase number N_c, and we use cross-entropy loss for this multi-class classification task:

$$\mathcal{L}_C = -\sum_{t=1}^{T} c_t \log(p_t). \tag{5}$$

For workflow anticipation, we follow [4, 5] to reformulate it to estimate remaining time. Ground truth is generated from recognition label as $r_t \in \mathbb{R}^{N_r}$, also with phase number. Each ranges [0,h] where 0 represents current occurrence, and h represents that the phase will not happen within next h minutes. We utilize smooth L1 loss for this regression task:

$$\mathcal{L}_R = -\sum_{t=1}^{T} \mathrm{SmoothL1}(f_t, r_t). \tag{6}$$

Overall training objective is the combination of both loss terms: $\mathcal{L} = \mathcal{L}^C + \lambda\mathcal{L}^R$, where λ is hyper-parameter to balance loss.

3 Experiments

3.1 Datasets and Evaluation Metrics

Datasets. We extensively evaluate our Trans-SVNet on three datasets, including two recording laparoscopic procedures, i.e., Cholec80 [7] and M2CAI16 Challenge dataset [24], and one recording cataract procedure from CATARACTS Challenge [25]. **i) Cholec80** consists of 80 cholecystectomy videos with 7 defined phases. Its frame resolution is either 1920×1080 or 854×480 and video is captured at 25fps. We follow the same evaluation protocol of previous works [7, 9, 10] using the first 40 videos for training and the rest for testing. **ii) M2CAI16** contains 41 videos that are segmented into 8 phases, divided into 27 for training and 14 for testing, following [9, 10]. Frame has 1920×1080 resolution and video is captured at 25fps. **iii) CATARACTS** (CATA.) consists of 50 videos obtained at 30 fps. Each frame is annotated with more detailed granularity of step, including 18 surgical steps and 1 idle step, with 1920×1080 resolution. We split the dataset into 25, 5, and 20 videos for training, validation and testing, following official challenge [25].

Evaluation Metrics. For workflow recognition, we employ four metrics for Cholec80 and M2CAI16, following previous works [7, 9, 14], including accuracy (AC), precision (PR), recall (RE), and Jaccard index (JA). We use F1 score to validate the performance of CATARACTS following challenge [25].

Table 1 Recognition results using different approaches on Cholec80 dataset.

Method	Accuracy	Precision	Recall	Jaccard (%) ↑	#param ↓
EndoNet* [7]	81.7 ± 4.2	73.7 ± 16.1	79.6 ± 7.9	-	58.3M
MTRCNet-CL* [19]	89.2 ± 7.6	86.9 ± 4.3	88.0 ± 6.9	-	29.0M
SV-RCNet [9]	85.3 ± 7.3	80.7 ± 7.0	83.5 ± 7.5	-	28.8M
OHFM [10]	87.3 ± 5.7	-	-	67.0 ± 13.3	47.1M
TeCNO [14]	88.6 ± 7.8	86.5 ± 7.0	87.6 ± 6.7	75.1 ± 6.9	24.7M
Trans-SVNet(Ours)	$\mathbf{90.9} \pm \mathbf{5.8}$	$\mathbf{91.4} \pm \mathbf{6.7}$	$\mathbf{89.4} \pm \mathbf{6.5}$	$\mathbf{79.7} \pm \mathbf{6.8}$	24.8M

Note: the $*$ means methods using extra tool labels.

Table 2 Recognition results using different approaches on M2CAI challenge dataset.

Method	Accuracy	Precision	Recall	Jaccard (%) ↑
SV-RCNet [9]	81.7 ± 8.1	81.0 ± 8.3	81.6 ± 7.2	65.4 ± 8.9
OHFM [10]	85.2 ± 7.5	-	-	68.8 ± 10.5
TeCNO [14]	86.1 ± 10.0	85.7 ± 7.7	88.9 ± 4.5	74.4 ± 7.2
Trans-SVNet(Ours)	$\mathbf{87.8} \pm \mathbf{8.0}$	$\mathbf{88.5} \pm \mathbf{5.3}$	$\mathbf{89.0} \pm \mathbf{4.9}$	$\mathbf{76.0} \pm \mathbf{7.9}$

For workflow anticipation, performance is evaluated by the variants of mean absolute error (MAE), including inMAE and eMAE, following [4, 5]. We evaluate the model on horizon h of 5 minutes for two laparoscopy datasets, and 1 minute for cataract dataset given its more fine-grained action annotations. We also count the number of model parameters and note that parameters are constant across different datasets.

3.2 Implementation Details

Our embedding and aggregation models are trained one after the other on PyTorch using an NVIDIA RTX 2080 GPU. For spatial embedding, we initialize ResNet parameters from pre-trained model on ImageNet [23]. It employs SGD with momentum of 0.9 and learning rate of 5e-4, except for the fully connected layers with 5e-5. Batch size is set to 100, and we resize frames into 250×250. Data augmentation is applied, including 224×224 cropping, random mirroring, and color jittering. For temporal embedding, we re-implement TeCNO [14] based on their released code with only phase labels and make the output of its second stage as our temporal embedding. We report the re-implemented results of TeCNO. The spatial and temporal embeddings extracted from ResNet50 and TeCNO are used as inputs to our aggregation model, without further tuning when we train aggregation model. Our aggregation model is trained by Adam with learning rate of 1e-3 and batch size identical to the length of each video. The number of attention heads is empirically set to 8, and the sequence length n is 30. N is the sum of dimensions of both recognition and anticipation labels, i.e., double of the phase/step number.

3.3 Comparison with State-of-the-arts

Workflow recognition. The compared methods for laparoscopy datasets, i.e., Cholec80 and M2CAI16, are relatively consistent, apart from several multi-task learning methods for Cholec80 using available tool labels. Therefore, we analyze their results together, shown in Table 1 and Table 2. We can see that using extra tool annotations of Cholec80, multi-task learning [7, 19] generally

Table 3 Recognition results of F1 score using different approaches on CATARACTS.

Results ↑	hutom-ARM	SK	Uniandes-BCV	Trans-SVNet(Ours)	CAMI-SIAT*	ARTG*
Validation	0.2497	0.5221	0.6419	**0.7082**	0.7115	**0.7890**
Test	-	0.8181	0.7827	**0.8402**	0.8242	**0.8920**

Note: the * means methods using extra labels.

Table 4 Anticipation results under different settings.

Results ↓	Cholec80 (60/20) with different methods			Ours on different datasets		
	BayesianDL[4]	IIA-Net*[5]	Trans-SVNet	Cholec80 (40/40)	M2CAI16	CATA.
inMAE	1.17	1.08	**1.07**	1.25	1.24	0.29
eMAE	1.37	**1.22**	1.26	1.53	1.38	0.32

Note: the * means the method leveraging other tasks for supervision.

achieve higher performances, where MTRCNet-CL surpasses all single-task models except ours. As for methods only using phase label, compared with multi-step learning OHFM, our approach attains a significant improvement by 6%-12% JA with a much simpler training. Compared with TeCNO with the same backbones, our Trans-SVNet gains a boost by 4% in PR and JA on the larger Cholec80 dataset with a negligible increase in model parameters. Note that our method achieves a more remarkable improvement on the larger Cholec80 than M2CAI16. The underlying reason is that the robustness of our method can yield a better advantage on a more complicated Cholec80 dataset.

We further validate on CATARACTS and compare with state-of-the-arts, i.e., top-five approaches in challenge. Among them, CAMI-SIAT utilizes additional phase annotations, and ARTG employs the available tool labels. Results are listed in Table 3. We can observe that our Trans-SVNet largely outperforms other approaches that solely use the surgical step labels. Our method even attains the better results over CAMI-SIAT that uses extra information. Although ARTG achieves better performance, the used tool information in fact cost the large annotation efforts in clinical practice.

Workflow anticipation. We also evaluate our Trans-SVNet on workflow anticipation task. Existing methods are relatively limited and only validated on Cholec80 with different dataset split (60 videos for training and 20 for testing). For fair comparison, we follow this split and compare our method with the state-of-the-arts in Table 4. [5] introduces extra feature extraction modules for detection and segmentation to assist anticipation, and requires additional training datasets. Our method with neat architecture achieves competitive results to [5]. We also shows our results on all three datasets with the same split as recognition task, e.g., 40/40 for Cholec80. Our method consistently attains promising performance on all datasets on different MAE variants. Compared with the results of the setting 60/20, no large performance degradation is shown even when we utilize much fewer labeled video for training (the setting 40/40). Regarding model efficiency, our method can generate predictions of two different tasks at ~ 0.0134 second with one GPU. It even largely exceeds the video recording speed.

Table 5 Phase recognition results of different architectures on Cholec80.

Architecture			Accuracy	Precision	Recall	Jaccard (%) ↑	P-values
PureNet	ResNet		82.1 ± 7.8	78.0 ± 6.4	78.5 ± 10.8	61.7 ± 11.3	2e-8
	TeCNO		88.6 ± 7.8	86.5 ± 7.0	87.6 ± 6.7	75.1 ± 6.9	2e-7
	ResNet *cat* TeCNO		87.9 ± 7.5	86.6 ± 5.9	85.3 ± 8.2	73.0 ± 7.8	2e-8
	Query	Key	Accuracy	Precision	Recall	Jaccard (%) ↑	P-values
Transformer	l_t	$l_{t-n+1:t}$	81.9 ± 9.2	78.0 ± 12.5	78.3 ± 12.8	60.8 ± 12.4	2e-8
	g_t	$g_{t-n+1:t}$	89.1 ± 7.8	87.6 ± 6.3	87.7 ± 6.9	76.2 ± 6.6	4e-7
	g_t	$l_{t-n+1:t}$	89.2 ± 7.5	87.7 ± 6.7	87.7 ± 7.0	76.1 ± 7.0	3e-7
	l_t	$g_{t-n+1:t}$	**90.3 ± 7.1**	**90.7 ± 5.0**	**88.8 ± 7.4**	**79.3 ± 6.6**	—

Table 6 Ablation study on multi-task learning for workflow recognition.

Methods	Accuracy	Precision	Recall	Jaccard (%) ↑	F1 Score ↑
	Cholec80				CATA. (Val.)
Single	90.3 ± 7.1	90.7 ± 5.0	88.8 ± 7.4	79.3 ± 6.8	0.6725
Multi	**90.9 ± 5.8**	**91.4 ± 6.7**	**89.4 ± 6.5**	**79.7 ± 6.8**	**0.7082**
	M2CAI16				CATA. (Test.)
Single	87.2 ± 9.3	88.0 ± 6.7	87.5 ± 5.5	74.7 ± 7.7	0.8259
Multi	**87.8 ± 8.0**	**88.5 ± 5.3**	**89.0 ± 4.9**	**76.0 ± 7.9**	**0.8402**

3.4 Analytical Ablation Study

We first study the impact of the way to incorporating hybrid embeddings, with recognition task on Cholec80. Models are trained toward single cross entropy loss for a clear and direct comparison. All results are listed in Table 5. Specifically, we first implement baselines without Transformer (PureNet), to validate the settings of solely using space (ResNet) or time information (TeCNO). We study an vanilla solution to consider both embeddings by concatenating l_t and g_t (ResNet *cat* TeCNO) and employing a linear layer to process. However, it is difficult to let one FC layer well adapt to the high-dimensional concatenated feature without end-to-end training the whole framework. Straightforward performing feature concatenation may also bring feature redundancy issue, causing parameter increase of the FC layer and increasing the risk of model overfitting to the training dataset. Therefore, the performance falls between ResNet and TeCNO. We also explore different configurations in Transformer, by varying embeddings from ResNet and TeCNO to be query or key in every possible combinations. We can see that only with the spatial embedding, i.e., l_t querying $l_{t-n+1:t}$, results even slightly decrease compared with pure ResNet. The reason is that spatial embeddings cannot indicate their orders in videos which are essential for Transformer-based feature aggregation. Superior results to PureNet are achieved after introducing temporal embeddings g to be query or key, justifying that Transformer rediscovers necessary details neglected by temporal extractors. We use l_t to query $g_{t-n+1:t}$ in Trans-SVNet which generates the best outcomes. We further conduct the statistical analysis on significance. We utilize wilcoxon signed-rank test to calculate P-values in JA for compared settings towards our Trans-SVNet. Apart from obvious increase in the results of JA, we found that P-values are substantially less than 0.05 in all compared cases, which indicates a significant improvement by using our proposed architecture.

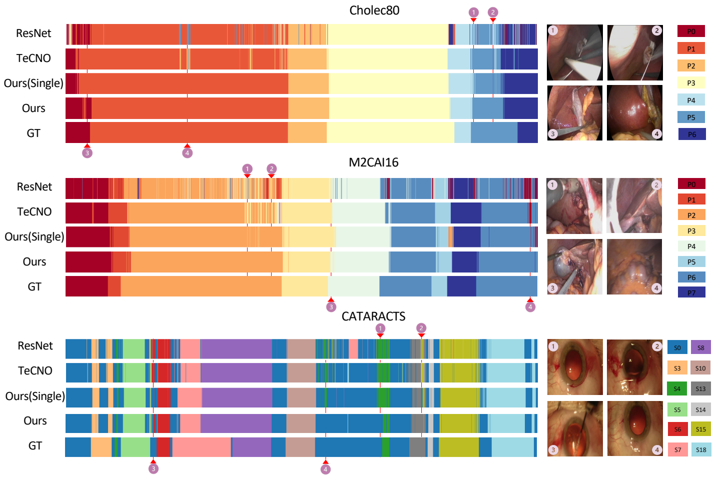


Fig. 3 Color-coded ribbon illustration for three complete surgical videos.

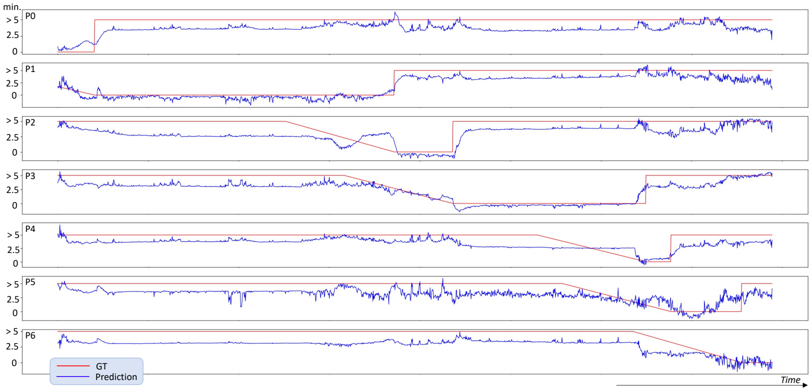


Fig. 4 Anticipation results of the remaining time (min.) until occurrence of each phase.

We then study the efficacy of introducing anticipation loss objectives to recognition task. It is observed from Table 6 that joint learning with anticipation, we receive the consistent performance increase across three datasets on recognition, indicating the complementary benefits brought from anticipation task. Results yield the greatest gain on the most complex cataract dataset.

3.5 Visual Results

We illustrate visual results of both tasks on the complete videos in Fig. 3 and Fig. 4. We choose the video sequences that can cover most surgical phases/steps, and present the typical workflow procedure in laparoscopy and cataract surgery. Fig. 3 shows recognition results via color-coded ribbon under four settings from all the three datasets. We can see that aggregating embeddings from ResNet and TeCNO wisely, our Trans-SVNet trained with single recognition task already contributes to more consistent and robust predictions, especially for Cholec80 P1, M2CAI16 P2, and CATA. S0. Our full model that jointly learns both tasks attains more complete results, such as Cholec80 P0, M2CAI16 P4, and CATA. S13. We visualize some typical frames for intuitive analysis, where we find that the mis-classifications are generally due to challenging scenes, that are negatively influenced by lighting reflection and motion blur. In Fig. 4, we show the anticipation results of all phases in a complete video from Cholec80. Our model is robust that predictions of remaining time in most phases are consistent to ground truths, especially for P1 and P3.

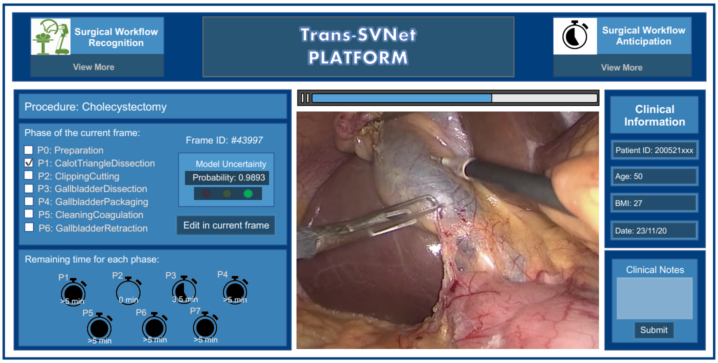
4 Discussion and Conclusion

Recognizing and anticipating workflow from surgical video are fundamentally crucial for various applications in CAI system. Previous methods fall into a conventional feature extraction routine. Our Trans-SVNet rethinks procedure, by using Transformer to explicitly consider hybrid features when calculating final representations. Our recognition results are consistently superior to state-of-the-arts across three datasets. We also achieve promising and competitive anticipation results, which shine a light and can inspire further investigations on using Transformer for this task.

Temporal relationships in surgical videos exhibit various behaviors, such as different phase duration, making it hard to define a universal temporal range to extract essential clues. Existing LSTM and TCN extractors maintain previous information with a fixed size of hidden states and include the latest cues with unchangeable procedure. Instead of a passive way for decision-making, we propose to leverage Transformer to actively select information from a embedding sequence, which alleviates side-effects of irrelevant ones. Moreover, recognizing some phases is highly dependent on visual appearances, such as Preparation phase. We exploit spatial embedding to query information, and leverage skip connection in Transformer, therefore, critical spatial information is preserved till the final decision stage.

Our model shows great potential usability in clinical practice from computational efficiency (14.3 ms per frame) and generalization (promising results on three datasets with two types). This is attributed to several reasons. Apart from parallel processing in Transformer, we found one interesting phenomenon that fusing features with very short lengths, e.g., 7 in our setting, already contributes to remarkable boosts. Such compact representations can extract critical information, enabling fast speed and improving results. Compact space may also facilitate generalization by alleviating overfitting. Additionally, our multi-task learning by introducing anticipation does not involve extra annotation effort or usage limitation, also bring few computational cost.

We identify directions of improvement for the future investigation. First, we are encouraged to develop a prototype platform, with interface design shown in Fig. 5 and we will integrate our method as two function modules. Thanks to our efficient speed, the platform allows real-time instruction for automated surgical coaching. It can also be potentially utilized during intra-operation,



Procedure: Cholecystectomy

Phase of the current frame:

Frame ID: *#43997*

- □ P0: Preparation
- ☑ P1: CalotTriangleDissection
- □ P2: ClippingCutting
- □ P3: GallbladderDissection
- □ P4: GallbladderPackaging
- □ P5: CleaningCoagulation
- □ P6: GallbladderRetraction

Model Uncertainty

Probability: 0.9893

Remaining time for each phase:

Clinical Information

Patient ID: 200521xxx

Age: 50

BMI: 27

Date: 23/11/20

Fig. 5 The protocol Trans-SVNet platform for user experience of surgeons.

providing real-time supports for surgeons. Additionally, the method can be evaluated with user experiments of surgeons, to receive practical feedback for methodology development. Second, we plan to leverage multiple signals that are widely available in operating theatres, such as multi-view, depth cues, and communication audio. Multi-modal learning can be developed, such as regarding different signals as query and key under Transformer architecture.

Acknowledgements. This work is supported by Hong Kong Research Grants Council TRS project T42-409/18-R and CUHK T Stone Robotics Institute, and the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z]; and Horizon 2020 FET (863146). For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission Robotics Institute.

5 Declarations

Conflict of interest. Authors declare that they have no conflict of interest.
Ethical approval. This article does not contain any studies with human participants or animals performed by any of the authors.
Informed consent. This articles does not contain patient data.

References

[1] Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S.: Surgical data science for next-generation interventions. Nature Biomedical Engineering (2017)

[2] Padoy, N.: Machine and deep learning for workflow recognition during surgery. Minimally Invasive Therapy & Allied Technologies **28**(2), 82–90 (2019)

[3] Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P.: Surgical data science–from concepts toward clinical translation. Medical image analysis **76**, 102306 (2022)

[4] Rivoir, D., Bodenstedt, S., Funke, I., Bechtolsheim, F.v., Distler, M., Weitz, J., Speidel, S.: Rethinking anticipation tasks: Uncertainty-aware anticipation of sparse surgical instrument usage for context-aware assistance. In: MICCAI, pp. 752–762 (2020). Springer

[5] Yuan, K., Holden, M., Gao, S., Lee, W.-S.: Surgical workflow anticipation using instrument interaction. In: MICCAI, pp. 615–625 (2021). Springer

[6] Forestier, G., Riffaud, L., Jannin, P.: Automatic phase prediction from low-level surgical activities. IJCARS **10**(6), 833–841 (2015)

[7] 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 TMI **36**(1), 86–97 (2017)

[8] Lalys, F., Bouget, D., Riffaud, L., Jannin, P.: Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures. IJCARS **8**(1), 39–49 (2013)

[9] 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 TMI **37**(5), 1114–1126 (2018)

[10] Yi, F., Jiang, T.: Hard frame detection and online mapping for surgical phase recognition. In: MICCAI (2019)

[11] Twinanda, A.P., Yengera, G., Mutter, D., Marescaux, J., Padoy, N.: Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations. IEEE TMI **38**(4), 1069–1078 (2018)

[12] Funke, I., Bodenstedt, S., Oehme, F., von Bechtolsheim, F., Weitz, J., Speidel, S.: Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. In: MICCAI (2019)

[13] Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: CVPR, pp. 156–165 (2017)

[14] 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 (2020)

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

[16] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A.: A survey on visual transformer. arXiv preprint arXiv:2012.12556 (2020)

[17] Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: ECCV, pp. 214–229 (2020). Springer

[18] Wang, Y., Solomon, J.M.: Deep closest point: Learning representations for point cloud registration. In: CVPR, pp. 3523–3532 (2019)

[19] Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.-W., Heng, P.-A.: Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Medical Image Analysis **59**, 101572 (2020)

[20] Zhang, J., Nie, Y., Lyu, Y., Yang, X., Chang, J., Zhang, J.J.: Sd-net: joint surgical gesture recognition and skill assessment. IJCARS **16**(10), 1675–1682 (2021)

[21] Franke, S., Neumuth, T.: Adaptive surgical process models for prediction of surgical work steps from surgical low-level activities. In: 6th Workshop on M2CAI at MICCAI (2015)

[22] 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: MICCAI, pp. 593–603 (2021). Springer

[23] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)

[24] Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: MICCAI M2CAI Challenge. http://camma.u-strasbg.fr/m2cai2016/

[25] Al Hajj, H., Lamard, M., Conze, P.-H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O.: Cataracts: Challenge on automatic tool annotation for cataract surgery. Medical image analysis **52**, 24–41 (2019)