Medical Image Analysis 105 (2025) 103724

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

A new dataset and versatile multi-task surgical workflow analysis framework for thoracoscopic mitral valvuloplasty

Meng Lan [a], Weixin Si [c], Xinjian Yan [b,*], Xiaomeng Li [a,*]

[a] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region of China
[b] Department of Cardiac Surgery, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, 510440, China
[c] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China

ARTICLE INFO

Keywords:
Surgical workflow analysis
Long-form video dataset
Multi-task learning

ABSTRACT

Surgical Workflow Analysis (SWA) on videos is critical for AI-assisted intelligent surgery. Existing SWA methods primarily focus on laparoscopic surgeries, while research on complex thoracoscopy-assisted cardiac surgery remains largely unexplored. In this paper, we introduce **TMVP-SurgVideo**, the first SWA video dataset for thoracoscopic cardiac mitral valvuloplasty (TMVP). TMVP-SurgVideo comprises 57 independent long-form surgical videos and over 429K annotated frames, covering four key tasks, namely phase and instrument recognitions, and phase and instrument anticipations. To achieve a comprehensive SWA system for TMVP and overcome the limitations of current SWA methods, we propose **SurgFormer**, the first query-based Transformer framework that simultaneously performs recognition and anticipation of surgical phases and instruments. SurgFormer uses four low-dimensional learnable task embeddings to independently decode representation embeddings for the predictions of the four tasks. During the decoding process, an information interaction module that contains the intra-frame task-level information interaction layer and the inter-frame temporal correlation learning layer is devised to operate on the task embeddings, enabling the information collaboration between tasks within each frame and temporal correlation learning of each task across frames. Besides, SurgFormer's unique architecture allows it to perform both offline and online inferences using a dynamic memory bank without model modification. Our proposed SurgFormer is evaluated on the TMVP-SurgVideo and existing Cholec80 datasets to demonstrate its effectiveness on SWA. The dataset and the code are available at https://github.com/xmed-lab/SurgFormer.

1. Introduction

Surgical workflow analysis, the core of surgical data science system (Maier-Hein et al., 2022a), is essential for intelligent computer-assisted surgery systems (Jin et al., 2022; Maier-Hein et al., 2022b). Such systems can monitor the surgical process, facilitate scheduling, and ultimately enhance patient safety (Maier-Hein et al., 2017). Thoracoscopy-assisted cardiac mitral valvuloplasty is a critical minimally invasive treatment for patients with mitral regurgitation, which is the most common form of heart valve disease (Otto et al., 2021). However, the complex structure of the heart, the need for precise coordination of surgical instruments, and the geometric constraints of the surgical field present significant challenges for novice surgeons during TMVP (Zhong and Huang, 2023; Hayashi et al., 2021). Unfortunately, existing SWA methods have primarily been designed and evaluated on laparoscopic or endoscopic videos, while the process of TMVP is more complex and poses a higher level of operational difficulty than laparoscopic surgeries, which may make it hard for existing SWA methods to analyze the workflow of TMVP effectively. Therefore, developing an accurate SWA framework for TMVP is crucial for monitoring the surgical process and ensuring surgical safety.

Since state-of-the-art SWA methods are data-driven deep learning models, we first collect a large-scale dataset TMVP-SurgVideo, consisting of 57 long-form videos of TMVP, to train reliable models for task executions. Typically, the surgical process models (Gibaud et al., 2018) decompose the surgical process into a multi-granular hierarchical structure, such as phase, step, and activity. Among them, phase recognition is a widely studied task in the SWA field. Additionally, researchers have introduced the instrument recognition task (Jin et al., 2020), as well as the phase and instrument anticipation tasks (Yuan et al., 2022; Jin et al., 2022) to provide context-aware assistance and improve surgical preparation efficiency. Therefore, to perform a comprehensive SWA

* Corresponding authors.
E-mail addresses: eexmli@ust.hk (X. Li), yanxinjian@gdph.org.cn (X. Yan).

https://doi.org/10.1016/j.media.2025.103724
Received 8 November 2024; Received in revised form 5 June 2025; Accepted 6 July 2025
Available online 21 July 2025
1361-8415/

system for TMVP, we design four task settings based on the collected dataset: phase recognition, instrument recognition, phase anticipation, and instrument anticipation. We also provide frame-level annotations for all four tasks. However, building separate models for each task to achieve comprehensive SWA is labor-intensive, and such single-task models often struggle to understand the complex scenes in TMVP-SurgVideo. Consequently, developing a versatile multi-task framework that can implement these tasks in a unified model and collaborate the information of surgical phases and instruments offers significant potential for achieving accurate surgical workflow perception.

Existing multi-task SWA methods primarily focus on two tasks (Twinanda et al., 2017; Ramesh et al., 2021), such as phase and instrument recognition (Wagner et al., 2023), and a few studies (Valderrama et al., 2022) explored the four-task framework. However, the existing multi-task SWA frameworks present two issues. **First**, most multi-task frameworks simply add task heads to conventional single-task architectures, inheriting their limitations. **Second**, existing multi-task architectures lack explicit inter-task collaboration, processing features like single-task models and hindering discriminative representation learning across tasks.

To address these problems, we propose SurgFormer, a straightforward yet effective multi-task SWA framework. Specifically, SurgFormer employs a query-based Transformer architecture, similar to Deformable DETR (Zhu et al., 2021), where four low-dimensional learnable task embeddings serve as the queries to decode task representations from the input image features in the decoder for final task predictions. An information interaction module is devised to facilitate task representation interaction and temporal correlation learning. Additionally, SurgFormer's architecture uniquely supports both online intraoperative analysis and offline postoperative video processing, establishing it as a versatile SWA system.

In summary, we conclude our contributions as follows:

(1) We introduce TMVP-SurgVideo, the first large-scale surgical video dataset for TMVP, with comprehensive per-frame annotations for four tasks. This dataset will significantly advance SWA research in complex cardiac procedures.
(2) We propose SurgFormer, the first query-based Transformer framework for versatile multi-task SWA. The model leverages low-dimensional task embeddings to learn representation for prediction of each task, which also enables SurgFormer to perform both online and offline analysis within a single model.
(3) We devise an information interaction module to realize explicit information collaboration among tasks and temporal correlation learning across frames.
(4) We evaluate the performance of various representative SWA methods on the TMVP-SurgVideo dataset, and the results demonstrate the challenging nature of TMVP-SurgVideo and the effectiveness of SurgFormer.

2. Related work

2.1. Surgical video datasets

As shown in Table 1, current public datasets for SWA primarily focus on laparoscopic surgeries, and phase and instrument recognitions are the fundamental tasks in SWA. However, due to the substantial annotation workload, the instrument-related labels of most datasets typically cover only a subset of all frames. For instance, the popular Cholec80 dataset (Twinanda et al., 2017) provides the phase the instrument recognition labels for all frames, while the HeiCo (Maier-Hein et al., 2021), PSI-AVA (Valderrama et al., 2022) and AutoLaparo (Wang et al., 2022) datasets release all phase annotations but only partial instrument labels, as shown in Table 1. The ESD dataset (Cao et al., 2023) focuses on phase recognition, creating a large-scale dataset with only phase annotations. The Bypass40 dataset (Ramesh et al., 2021) collects 40 surgical videos annotated with both phases and steps. Different from these existing datasets, this paper focuses on the complex TMVP. We introduce the first large-scale TMVP dataset with comprehensive annotations for recognition and anticipation of both phase and instrument.

2.2. SWA methods

Single-task Methods The rapid advancement of computer vision has driven the development of vision-only SWA systems. Typically, the complex surgical workflow can be decomposed into hierarchical structures, so as to provide operators with surgical information at multiple levels of granularity (Lalys and Jannin, 2014).

Phase recognition, a major and essential task in the SWA field, has been widely studied by researchers for readily available annotations. In the early era, some works (Twinanda et al., 2017; Blum et al., 2010; Quellec et al., 2014) used linear statistical models to capture the temporal information of videos. However, their performance is limited by the empirically designed features. The introduction of the LSTM (Graves and Graves, 2012) improved temporal correlation learning of videos in Jin et al. (2018) and Yi and Jiang (2019) but faced limitations in capturing long-term dependencies due to the vanishing gradient problem. Inspired by TCNs (Farha and Gall, 2019; Czempiel et al., 2020) proposed the TeCNO for learning long-term temporal correlations. Ding and Li (2022) employed the cascaded TCNs to explore the segment-level semantics and refine erroneous predictions. Further, Ding et al. (2023) utilized the timestamp as weak supervision to train a TCN-based model for surgical phase recognition. However, the inherent properties of dilated convolutions in TCNs can result in information loss in long sequences. The emergence of attention mechanism (Vaswani et al., 2017), renowned for its ability to establish long-range dependencies, led to the adoption of the Transformer architecture for phase recognition (Gao et al., 2021; Liu et al., 2023; Yue et al., 2023). For instance, Gao et al. (2021) devised a hybrid embedding aggregation transformer to fuse the spatial and temporal features. Czempiel et al. (2021) integrated multiple self-attention layers behind a CNN to create temporal relationships among frame features. Liu et al. (2023) developed the key information Video Transformer to record global key information along the temporal dimension to guide the prediction. However, the quadratic time and memory complexity caused by attention operations on dense spatio-temporal features remain a challenge for long videos.

Since step recognition can be regarded as fine-grained phase recognition, all the phase recognition methods mentioned above can be adapted to the step recognition task (Shah et al., 2023). For surgical activity recognition, researchers usually treat each surgical activity as an action triplet <instrument, verb, target>. In Nwoye et al. (2020), Nwoye et al. built a relevant dataset and proposed a weakly supervised detection method to recognize surgical triplets. Nwoye et al. (2022) proposed the Rendezvous which introduced spatial attention and semantic attention to capture the action triplets. Later Sharma et al. (2023) extended the Rendezvous model with temporal modeling to better integrate the current and past features. Chen et al. (2023) devised a triplet disentanglement framework that decomposes the learning objectives to reduce learning difficulties.

Multi-task Methods. To fully leverage the multi-granularity information present in surgical videos, a number of multi-task frameworks have been proposed. Twinanda et al. (2017) introduced EndoNet, a multi-task framework for phase and tool recognition, but it lacked the incorporation of critical temporal dependencies in the model. Jin et al. (2020) developed a multi-task recurrent convolutional network for the phase and instrument recognition tasks and utilized a correlation loss to boost the performance of both tasks. In Wagner et al. (2023), the researchers concluded the multi-task SWA approaches proposed by the participants in the EndoVis Challenge 2019. The majority of these methods focused on phase and instrument recognition tasks, with most employing ResNet-50 as the backbone network and utilizing

Table 1
The statistics comparison of existing SWA datasets and our TMVP-SurgVideo dataset. The transition phase means that phase 0 acts as a transition between other adjacent phases, which are not seamlessly connected, as shown in Fig. 1(a).

Dataset	Video number	Frames	Surgery types	Number of annotated phases and instruments	Tasks (number of annotated frame)
Cholec80 (Twinanda et al., 2017)	80	184,578 (1 fps)	Laparoscopy-assisted surgery	7 phases without transition phase, 7 instruments	Phase recognition (184,578 frames), instrument recognition (184,578 frames)
PSI-AVA (Valderrama et al., 2022)	8	73,618 (1 fps)		11 phases without transition phase, 7 instruments	Phase recognition (73,618 frames), instrument detection (2238 frames)
AutoLaparo (Wang et al., 2022)	21	83,280 (1 fps)		7 phases without transition phase, 4 instruments	Phase recognition (83,280 frames), instrument segmentation (1800 frames)
Bypass40 (Ramesh et al., 2021)	40	256,000 (1 fps)		11 phases without transition phase, 0 instrument	Phase recognition (256,000 frames)
HeiCo (Maier-Hein et al., 2021)	30	346,032 (1 fps)		14 phases without transition phase, 5 instruments	Phase recognition (346,032 frames), instrument segmentation (10,040 frames)
ESD (Cao et al., 2023)	47	201,026 (1 fps)	Endoscopy-assisted surgery	4 phase without transition phase, 0 instrument	Phase recognition (201,026 frames)
TMVP-SurgVideo	57	**429,494** (1 fps)	Thoracoscopy-**assisted cardiac surgery**	**12 phases with transition phase**, 8 instruments	Phase recognition, instrument recognition, **phase anticipation, instrument anticipation** (all these tasks are with 429,494 frames)

LSTM networks to aggregate temporal information. In Yuan et al. (2022), Yuan et al. proposed a two-stage model for the phase and instrument anticipation tasks, where the spatial features of phase and instrument were first extracted and then fed into a MS-TCN (Farha and Gall, 2019) for temporal pattern modeling. Subsequently, Jin et al. (2022) combined phase recognition and phase anticipation tasks within a two-stage Transformer framework and realized promising results. Valderrama (Valderrama et al., 2022) used a video feature extractor followed by a task-specific classification head to implement the phase and step recognitions and incorporated box-specific features extracted by an instrument detector to perform instrument-related tasks.

However, most existing multi-task methods typically adopt the traditional single-task architecture and ignore explicit information interaction among tasks, thus limiting their potential. In contrast, our query-based Transformer architecture introduces low-dimensional task embeddings for each task prediction and an information interaction module to enable efficient multi-task collaboration and temporal learning at minimal cost.

3. TMVP-SurgVideo dataset

3.1. Dataset statistics and analysis

TMVP-SurgVideo contains 57 distinct surgical videos from 57 patients operated by three expert surgeons and one novice surgeon. The number of cases performed by the three experts and one novice is 11, 14, 17, and 15, respectively. Each video in the dataset captures the whole TMVP procedure. All these videos were recorded using a thoracoscopic camera at 25 fps and a high resolution of 1280 × 720 pixels. To facilitate analysis, the videos were downsampled to 1 fps, resulting in a total of 429,494 frames. The total video duration of the dataset is 119.3 h, with a minimum video duration of 43.78 min, a maximum duration of 300.92 min, and a standard deviation of 39.96 min. TMVP-SurgVideo is divided into a training set (30 videos) and a validation set (27 videos). In both training and validation sets, we keep the ratio of the number of surgical videos operated by experts and novice surgeons at about 3:1. Based on the TMVP-SurgVideo dataset, we design four tasks, i.e., phase recognition, phase anticipation, instrument recognition, and instrument anticipation. Under the guidance of expert surgeons, each frame in the dataset has annotations of the four tasks.

Advantage. As shown in Table 1, compared to previous video datasets on laparoscopic or endoscopic surgery, (1) TMVP-SurgVideo is the first large-scale dataset for thoracoscopy cardiac mitral valvuloplasty, expanding the scope of SWA research to encompass complex cardiac TMVP and enriching the diversity of available data. (2) TMVP-SurgVideo provides complete annotations of the four tasks for all 429,494 frames, allowing researchers to explore both single-task and multi-task SWA approaches.

Challenges. (1) Unlike previous datasets where phases seamlessly transition, TMVP-SurgVideo includes a preparation phase (P0) that may occur between any adjacent phases as transition, as shown in Fig. 1(a), which increases the complexity to phase recognition and anticipation. (2) The dataset exhibits significant intra-phase variance and inter-phase similarity, as presented in Fig. 2, increasing the risk of phase misclassification. (3) Both phase and instrument annotations exhibit significant class imbalance, potentially leading to model overfitting on frequent categories and poor performance on rare categories. (4) The surgical videos in the TMVP-SurgVideo dataset present many complex surgical scenes, including smoke, motion blur, visual occlusion, blood leaking, tilted camera perspective, and unpredictable scene switchings, as depicted in Fig. 3. These complex scenes increase the analysis difficulty.

3.2. Recognition task

Task Description. Phase recognition aims to segment surgical videos into distinct, predefined phases, and each phase contains a sequence of video frames. In contrast, instrument recognition focuses on identifying predefined instrument categories present in each frame. Phase recognition constitutes a multi-class classification task, while instrument recognition is a multi-label classification problem. Notably, since phase recognition divides the procedure into different phases through visual cues and the usage of different instruments, it is possible to improve the performance of both phase and instrument recognition by collaborating the surgical phase and instrument information.

Phase recognition. At the guidance of technical guidelines of TMVP (Committee, 2023), two expert surgeons divide the surgical workflow of TMVP into 12 phases: **Phase 0** (P0): Preparation, **Phase 1** (P1): Suspend pericardium, **Phase 2** (P2): Dissociate vein, **Phase 3**(P3): Suture spacer, **Phase 4** (P4): Insert perfusion needle, **Phase 5** (P5): Block aorta, **Phase 6** (P6): Insert a left heart drain, **Phase**

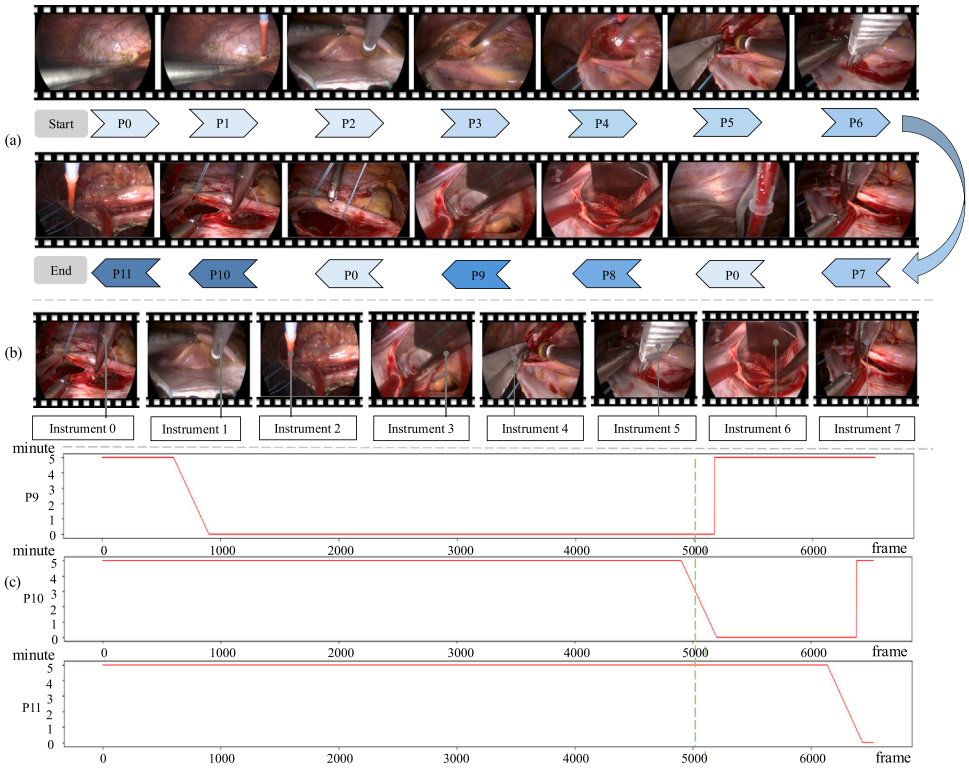

Fig. 1. Overview of TMVP-SurgVideo. (a) Order information of the procedure with sample frames of each phase. **Notably, P0 may occur between any two phases as a transition.** (b) Examples of the annotated instruments. (c) Illustrations of annotations of phase anticipation from P9 to P11. The dotted line means that at the 5000th frame, the current phase is 9, with approximately 3 min remaining until the transition to phase 10, and phase 11 is outside the threshold.

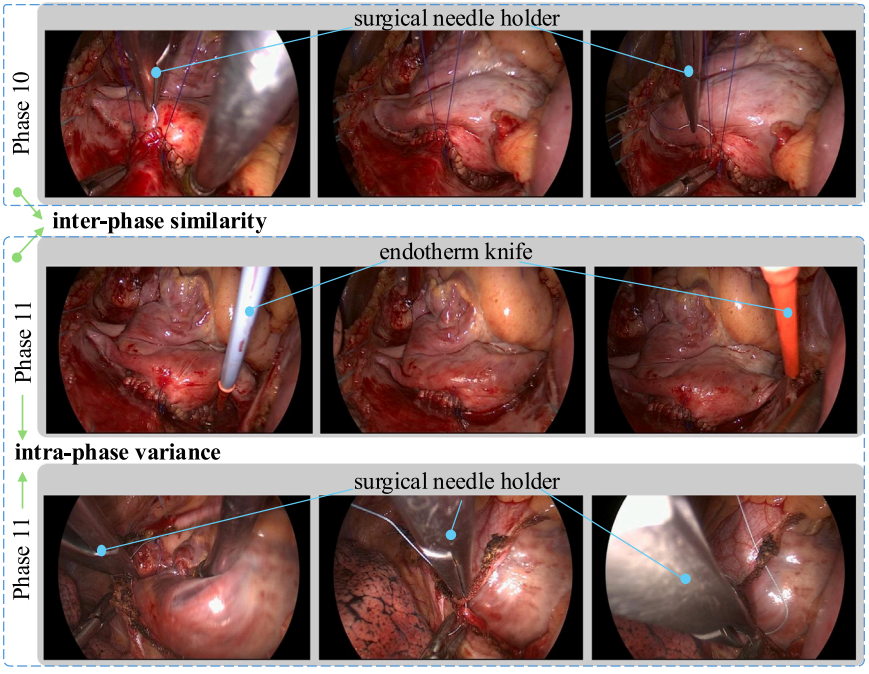

Fig. 2. Illustration of the inter-phase similarity and intra-phase variance in TMVP-SurgVideo.

7 (P7): Dissect the left atrium, **Phase 8** (P8): Expose mitral valve, **Phase 9** (P9): Mitral ValvuloPlasty, **Phase 10** (P10): Suture the left atrium and **Phase 11** (P11): Hemostasis and pericardium suture. P1 to P11 come in sequence, and the start and end of each phase from P1 to P11 are determined by the entry or removal of specific types of instruments, which results in them not appearing seamlessly, thus we insert P0 as a transition between adjacent phases that are not seamlessly connected. The number of annotated frames for each phase is successively [12549, 20193, 5779, 11547, 4573, 1375, 2783, 9494, 6690, 249730, 68078, 36703]. Compared to other public surgical video datasets, TMVP-SurgVideo exhibits more pronounced class imbalances, as illustrated in Fig. 4. Image examples of these phases are shown in Fig. 1(a). The experts provide start and end timestamps for each phase

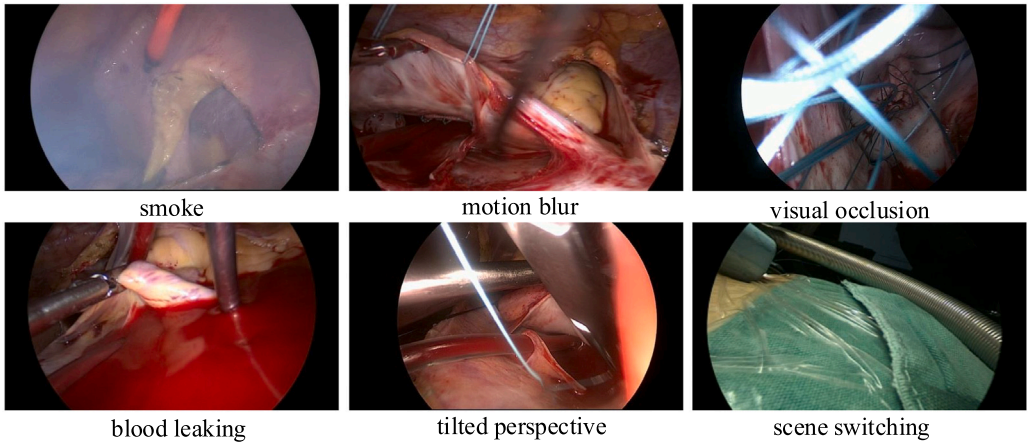

Fig. 3. Illustration of various complex scenes in TMVP-SurgVideo.

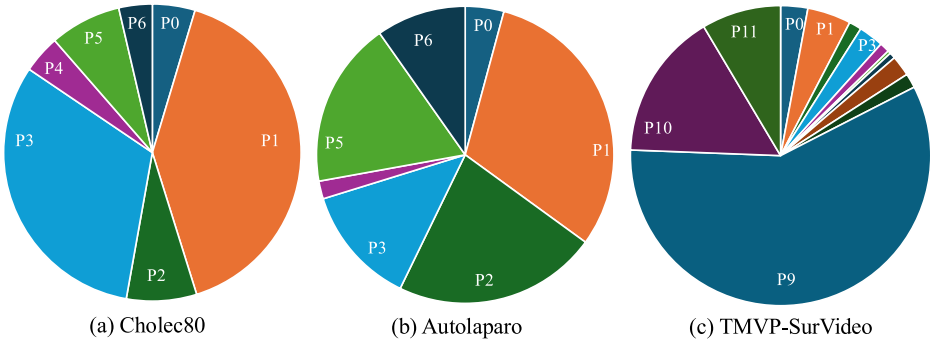

Fig. 4. Distribution of the number of frames for surgical phases in (a) Cholec80, (b) Autolaparo, and (c) TMVP-SurgVideo datasets.

of each video, and then a data engineer converts them into frame-level annotations.

Instrument Recognition. For the annotation of instruments used in the TMVP, under the suggestion of the expert surgeons, we focus on 8 representative instruments of the procedure, which are closely correlated with specific surgical phases and can provide discriminative information on instrument usage to improve the accuracy of both phase recognition and anticipation in the multi-task framework. The annotated instruments are as follows: **instrument 0:** surgical needle holder, **instrument 1:** aspirator, **instrument 2:** endotherm knife, **instrument 3:** thoracoscopic needle holder, **instrument 4:** occlusion forceps, **instrument 5:** knife, **instrument 6:** atrial retractor, **instrument 7:** surgical scissor. The number of annotated frames that contain each instrument is successively [36799, 42284, 15491, 78493, 2079, 2827, 153702, 9211]. The examples of the instruments are shown in Fig. 1(b). Since some video frames do not contain instruments that need to be recognized, the instrument annotations for these frames are set to 0, which are also used in the training and inference process. The number of these frames is 151857. During the annotation process, we adopt a semi-automatic annotation strategy. A data engineer first annotates 5 videos, approximately 34,000 images, with instrument annotations under the guidance of an expert surgeon of TMVP, then we train a multi-label instrument recognition model to automatically annotate the remaining 52 videos. Upon completion of the automatic annotation, the data engineer manually verifies the instrument annotations of all 57 videos, ensuring all the annotations are right. For additional quality assurance, the expert surgeon then randomly reviews around 10% of the annotations of each video as a final validation step. When there is dissensus on the annotation, the data engineers will discuss it with the expert surgeon to reach a consensus.

3.3. Anticipation task

Task Description. This task requires the analysis of real-time or recorded video data to foresee the sequence of events based on current and past surgical activities. By understanding the typical flow of TMVP and analyzing the surgeon's operations and instrument usage scenarios, the model can predict the time until the start of the subsequent phases or the start of other instrument usage. This anticipation ability allows for proactive preparation and response, ultimately aiming to enhance the efficiency and safety of surgical procedures.

We obtain the phase and instrument anticipation annotations based on the phase and instrument recognition annotations, respectively, as the phase and instrument recognition annotations are annotated in order and contain temporal information. For both phase and instrument anticipation, we set the anticipation horizon threshold h=5 min. This threshold indicates that predictions are made only for surgical phases or instruments that are anticipated to commence within the next 5 min (300 frames). Predictions for events occurring beyond this threshold are uniformly set to 5 min, and the anticipation prediction is set to zero for the current phase.

4. Methods

4.1. Overview

The overview of our proposed SurgFormer is presented in Fig. 5. SurgFormer adopts the query-based Transformer architecture for multi-task surgical workflow analysis. It takes a video sequence as input and outputs predictions for four tasks: phase recognition, tool recognition, phase anticipation, and tool anticipation. The model comprises four key components: image and Transformer encoders, a Transformer decoder, an information interaction module embedded in the decoder,

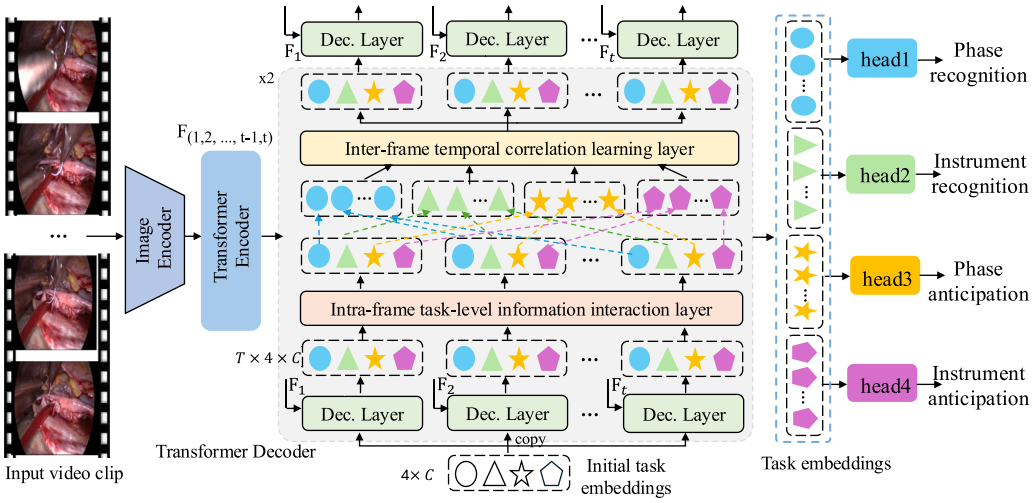

Fig. 5. An overview of the proposed SurgFormer. It mainly consists of four parts: the image and Transformer encoders, the Transformer decoder, the information interaction module embedded in the decoder, and the task-specific prediction heads for final predictions. In the Transformer decoder, four task embeddings interact independently with the input frame features to produce task-specific representation embeddings. These embeddings are then passed to the information interaction module, where the intra-frame task-level information interaction layer and inter-frame temporal correlation learning layer facilitate the interaction of task representations within each frame and the learning of temporal correlations across all frames, respectively. After multiple iterations, the representation embedding for each task and frame is fed through a task-specific prediction head to generate the final prediction. Here one shape in task embeddings indicates one task.

and prediction heads for the final results. Notably, during the online inference, SurgFormer employs a dynamic memory bank that stores the previous task embeddings to guide the prediction of the current frame. To handle the scale changes of instruments and achieve efficient attention computation, SurgFormer adopts the Deformable DETR as the Transformer architecture (Zhu et al., 2021).

During inference, given a video sequence $\mathcal{V} = \{I_t\}_{t=1}^{T}$ with T frames, the image encoder first extracts the image features, which are then sent to the Transformer encoder to enhance the feature representation with long-range dependence. The Transformer decoder receives the enhanced image features along with four randomly initialized task embeddings. Through iterative decoding, it generates task-specific representation embeddings for each frame. A key component of the decoder is the information interaction module, which facilitates information collaboration between tasks in each frame and temporal correlation modeling. The module has two units and each comprises two layers: the intra-frame task-level information interaction layer, which enables information interaction between different tasks within each frame, and the inter-frame temporal correlation learning layer, which captures the temporal relationships across different frames of each task. Finally, the task-specific representation embeddings are fed into separate prediction heads, producing predictions for each frame and task.

4.2. Feature encoding

4.2.1. Image encoder

For the images in the input video sequence, we employ ResNet50 (He et al., 2016) as the image encoder to extract multi-scale visual features of each frame independently, resulting in the visual feature sequence $F_v = \{F_t'\}_{t=1}^{T}$, where F_t' denotes the multi-scale features for the tth frame. Specifically, F_t' consists of four levels of pyramid features, with the first three levels corresponding to the last three stage features of the image encoder, employing spatial strides of $\{8, 16, 32\}$. The final level is obtained by downsampling the 32-stride feature using a convolutional layer with a stride of 2, yielding a four-level pyramid feature with strides of $\{8, 16, 32, 64\}$.

4.2.2. Transformer encoder

A three-layer deformable Transformer encoder (Zhu et al., 2021) is built to enhance the long-range representation of the input features. Before feeding the multi-scale features into the Transformer encoder, they are projected to a uniform channel dimension $C = 256$, followed by the addition of a fixed 2D positional encoding to features of each frame to reinforce positional information. Subsequently, the encoder processes these multi-scale features using multi-scale deformable self-attention modules (Zhu et al., 2021) in a frame-independent manner. The output multi-scale features $F_E = \{F_t\}_{t=1}^{T}$ are then sent to the decoder.

4.3. Feature decoding

The core component of our multi-task framework is the feature decoding, which consists of the deformable Transformer decoder layers and the information interaction module. The decoding process not only generates task-specific representation embeddings for final predictions but also realizes information collaboration between tasks and temporal correlation information for better performance.

4.3.1. Transformer decoder

As depicted in Fig. 5, the Transformer decoder takes as input the image features F_E and four randomly initialized task embeddings $Q^0 \in R^{4\times C}$. The task embeddings are first copied T times to ensure that each frame of the input features is associated with four task embeddings. Subsequently, these embeddings engage in iterative interactions with the multi-scale image features for three iterations, aiming to decode the representation information of each task. This process produces a set of $N_q = 4T$ task embeddings for final predictions.

While the decoding process between task embeddings and image features is implemented using multi-scale deformable cross-attention modules within the decoder layers in a frame-independent manner, it results in a deficiency in inter-task communication of representation information within the frame and insufficient temporal correlation of the same task representation across frames. These limitations are particularly detrimental to the effectiveness of a multi-task SWA framework.

To mitigate these issues efficiently, we devise an information interaction module and integrate it between the decoder layers of the Transformer decoder. This module enables both intra- and inter-frame information interactions to be performed solely on low-dimensional representation embeddings.

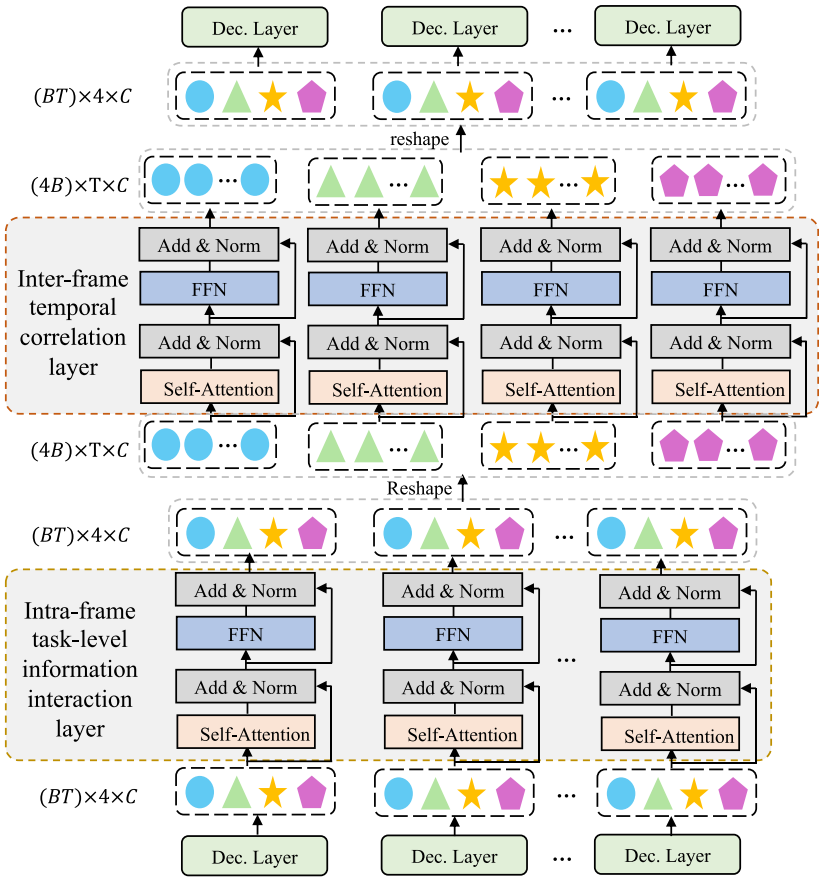

Fig. 6. The structure of one unit of the information interaction module. The i th unit of the module is embedded between the ith and the i+1th level of the decoder layers.

4.3.2. *Information interaction module*

As illustrated in Figs. 5 and 6, the information interaction module is composed of two units, each of them contains an intra-frame task-level information interaction (TLII) layer followed by an inter-frame temporal correlation learning (TCL) layer. The TLII layer aims to enhance the correlations among all task representations within each frame, while the TCL layer seeks to learn the temporal correlation of the same task representations across frames. Through this joint learning in the spatio-temporal dimension, more discriminative and robust task representations are generated for final predictions. Each unit of the module is embedded between two decoder layers, and the entire information interaction module is lightweight and plug-and-play for the decoding process.

Specifically, as shown in Fig. 6, after the decoding between the task embeddings and the image features in the decoder layer, the output representation embeddings $Q \in R^{(BT)\times 4\times C}$ are fed into the intra-frame TLII layer. Here, B is the batch size, T denotes the frame number of the input video sequence, and C is the dimension of the task embeddings. The four task embeddings of each frame independently pass through a multi-head self-attention layer and a Feed-forward Network (FFN) layer (Vaswani et al., 2017), which facilitate the interaction of representation information among the four tasks within the frame and enhance the correlation between task embeddings.

Following the intra-frame TLII interaction, the output task embeddings $Q_{intra} \in R^{(BT)\times 4\times C}$ are first reshaped to $Q_{intra} \in R^{(4B)\times T\times C}$ and then fed into the inter-frame TCL layer. In this layer, the embeddings of the same task from all the frames are grouped together, and the four resulting sequences are independently processed through a multi-head self-attention layer and a FFN layer. This facilitates the interaction of embeddings belonging to the same task within each sequence, fostering temporal correlation among the video frames. After the inter-frame interaction, the task embeddings are reshaped back to $Q_{inter} \in R^{(BT)\times 4\times C}$ for the subsequent decoding process in the decoder layer.

These two layers could be formulated as follows:

$$\begin{aligned} Q_1 &= LN(MHSA(Q) + Q), \\ Q_{intra} &= LN(FFN(Q_1) + Q_1), \\ Q_2 &= LN(MHSA(Re(Q_{intra})) + Re(Q_{intra})), \\ Q_{inter} &= Re(LN(FFN(Q_2) + Q_2)), \end{aligned} \tag{1}$$

where $Re(\cdot)$ indicates the reshape operation, $MHSA(\cdot)$ represents the multi-head self-attention layer (Vaswani et al., 2017), $LN(\cdot)$ denotes the layer normalization, and $FFN(\cdot)$ is the FFN layer.

Here, we analyze the complexity of the information interaction module. The complexity of the intra-frame TLII layer and the inter-frame TCL layer is $\mathcal{O}\left(4C^2 + 4^2C\right)$ and $\mathcal{O}\left(TC^2 + T^2C\right)$, respectively. Since C is constant and T is constrained to a constant due to the hardware limitation, the computation needed for the information interaction module almost could be neglected.

4.4. *Multi-task predictions*

After the decoding process in the Transformer decoder, $4T$ task embeddings with discriminative representations will be used to produce four final predictions of the four SWA tasks. Four prediction heads are employed, one for each task prediction, with each head implemented using a linear layer. The task embeddings for the same task of the input T frames are sent into the corresponding head to produce the final prediction results of those frames. To train the model, we calculate a loss for each task and use the total loss to optimize the model

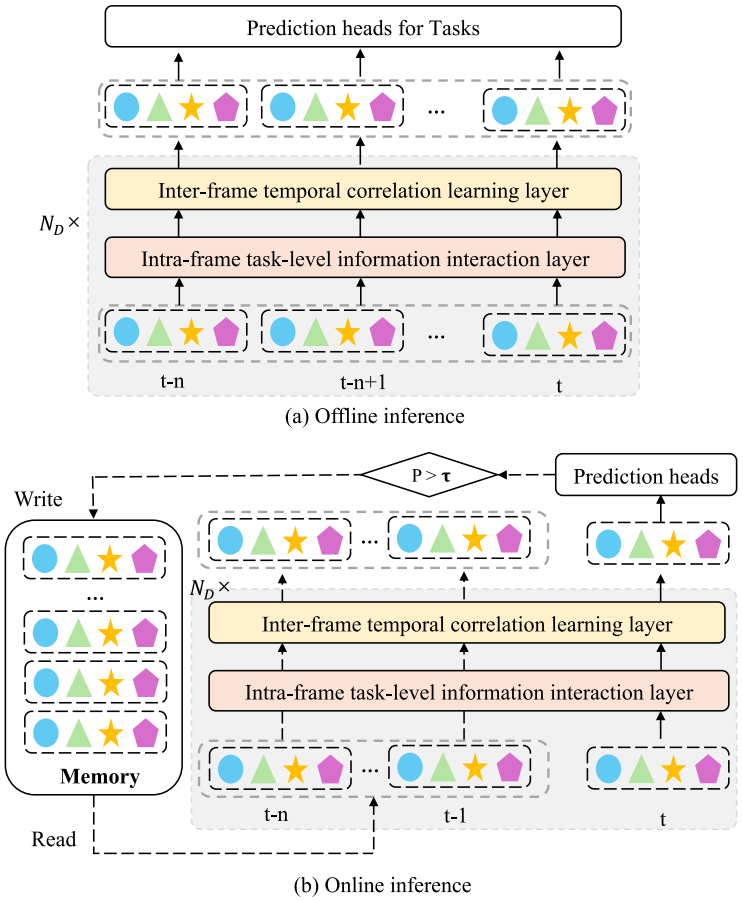

Fig. 7. The illustration of the decoding process for (a) offline inference and (b) online inference. The decoder layer is omitted for simplicity.

parameters. The loss function of the model is formulated as follows:

$$L\left(\left(p_1,p_2,p_3,p_4\right),\left(y_1,y_2,y_3,y_4\right)\right)=\lambda_1 L_{phase}\left(p_1,y_1\right) + \lambda_2 L_{ins}\left(p_2,y_2\right)+\lambda_3 L_{pha\_ant}\left(p_3,y_3\right)+\lambda_4 L_{ins\_ant}\left(p_4,y_4\right) \tag{2}$$

here, (p_1,p_2,p_3,p_4) and (y_1,y_2,y_3,y_4) are the predictions and ground truth of the four tasks, respectively. L_{phase} is the focal loss (Lin et al., 2017) for phase recognition, L_{ins} is the multi-label binary cross-entropy (BCE) logistic loss for instrument recognition, L_{pha_ant} and L_{ins_ant} are both the smooth L1 loss for phase anticipation and instrument anticipation, respectively. λ_1, λ_2, λ_3, and λ_4 are the loss weights of the corresponding loss functions, allowing the losses to be relatively balanced across tasks. Here, we chose focal loss to address the class imbalance in phase labels and adopt the BCE loss for a fair comparison with prior works (Jin et al., 2020).

4.5. Dual-mode inference

Prior studies have primarily focused on single-mode inference, specifically online inference, thereby limiting the model's applicability to postoperative scenarios. In contrast, our proposed Surgformer is capable of achieving both online and offline inference without changing the model architecture. The sole distinction between online and offline inference lies in the decoding process.

4.5.1. Offline-mode inference

As presented in Fig. 7(a), during offline inference, the model receives N video frames as input. After feature encoding, the decoder utilizes $4N$ task embeddings (four task embeddings per frame) to decode the N image features and produce $4N$ representation embeddings for the four-task predictions of the N frames. Throughout the decoding process, the $4N$ task embeddings are collectively processed by the information interaction module, generating discriminative task representation embeddings.

4.5.2. Online-mode inference

Different from offline inference, online-mode inference processes the video sequence frame-by-frame, starting with the first frame. To leverage information from previous frames, a memory bank is established to store a fixed number of task embeddings from processed frames. As depicted in Fig. 7(b), for the current frame t, after feature encoding, four task embeddings are employed to decode the task information from the features of frame t in the decoder layer. The resulting four task embeddings are then processed through the intra-frame TLII layer alone. Subsequently, the task embeddings corresponding to the same unit from previous frames in the memory bank are retrieved and fed into the inter-frame TCL layer alongside the four task embeddings from the current frame. Within the inter-frame TCL layer, the task embeddings of the current frame can leverage the task representation information of previous frames to enhance their own task representations. Upon obtaining the final prediction of each task, if the probability of phase recognition exceeds a threshold τ, the task embeddings at each level of the current frame will be saved in the memory bank, otherwise, they are discarded.

To ensure inference efficiency, the memory size is fixed, storing only the task embeddings of the preceding 64 frames. This memory adheres to a first-in-first-out rule: when the memory is full, if the task

embeddings of the current frame are to be stored, the task embeddings of the earliest incoming frame in the memory are removed.

5. Experiments

5.1. Datasets and evaluation metrics

Dataset. We evaluate the performance of several representative SWA methods on our TMVP-SurgVideo dataset and also compare our proposed SurgFormer with existing state-of-the-art methods on the Cholec80 dataset (Twinanda et al., 2017). The cholec80 dataset contains 80 videos of cholecystectomy surgeries performed by 13 surgeons. All the videos are recorded at 25 fps and have a resolution of 1920 × 1080 or 854 × 480. The dataset is divided into two subsets of equal size, with the first 40 videos serving as the training set and the remaining 40 as the testing set. To validate the generalizability of the proposed SurgFormer, we also collected and annotated an additional 12 surgical videos of TMVP procedures as an external validation set. The external validation set has a total of 80,299 frames, with an average duration of 111.5 min per video, and contains frame-level annotations for all four tasks. These cases were performed by three surgeons (two experts and one novice), with each surgeon contributing four cases. Notably, none of these surgeons' surgical videos are included in the TMVP-SurgVideo dataset.

Evaluation Metrics. Following the unrelaxed evaluation protocol in Liu et al. (2023), we evaluate the performance of SWA models across various metrics. For phase recognition, we employ three widely used metrics, following previous works (Jin et al., 2018; Czempiel et al., 2020), including precision (PR), recall (RE), and Jaccard (JA). PR, RE, and JA are used to validate the results at the phase level. We first compute PR, RE and JA for each phase using the following formulas: PR = $\frac{|GT \cap P|}{|P|}$, RE = $\frac{|GT \cap P|}{|GT|}$ and JA = $\frac{|GT \cap P|}{|GT \cup P|}$, where GT and P represent the ground truth and predictions, respectively. We then average these values across all phases to the overall PR, RE, and JA for the entire video. For tool recognition, a multi-class classification task, performance is evaluated using mean average precision (mAP), which accounts for both precision and ranking quality. We first calculate the AP values for each tool based on its precision–recall curve and then average these AP values across all the tools. For workflow anticipation, we adopt the frame-based metrics used in Rivoir et al. (2020) and Yuan et al. (2022) and adopt three variants of mean absolute error (MAE), i.e., inMAE, eMAE and pMAE, to access the performance. The three MAE variants can be represented as follows:

$$
\begin{aligned}
inMAE &= \frac{1}{T}\sum_{i}^{T} MAE\left(f_i, \alpha_i\right), 0 < \alpha_i < h \\
eMAE &= \frac{1}{T}\sum_{i}^{T} MAE\left(f_i, \alpha_i\right), 0 < \alpha_i < 0.1h \\
pMAE &= \frac{1}{T}\sum_{i}^{T} MAE\left(f_i, \alpha_i\right), 0.1h < f_i < 0.9h
\end{aligned}
\tag{3}
$$

where f_i is the prediction of the model, α_i is the ground truth value for the current timestamp, and h is the horizon threshold. Specifically, we average the MAE of 'anticipating' frames using inMAE, as the preparation system should only respond to the signals that indicate an impending phase/instrument change. eMAE reflects the anticipation performance within the nearest time period, providing the most effective support for the computer-assistance system. pMAE, on the other hand, indicates the precision of the predictions by considering only those frames with $0.1h < f_i < 0.9h$.

Table 2
Comparison results of phase recognition on TMVP-SurgVideo validation set.

Method	Mode	Backbone	PR	RE	JA
SV-RCNet (Jin et al., 2018)	Online	ResNet50	71.3	68.7	59.0
TeCNO (Czempiel et al., 2020)	Online	ResNet50	73.6	72.9	61.2
TMRNet (Jin et al., 2021)	Online	ResNet50	73.7	72.4	60.9
Trans-SVNet[a] (Jin et al., 2022)	Online	ResNet50	74.0	71.4	61.5
DINOv2[a] (Oquab et al., 2024)	Offline	ViT-B	74.5	72.8	62.4
MTRCNet[a] (Jin et al., 2020)	Online	ResNet50	74.9	73.2	62.5
SAHC (Ding and Li, 2022)	Online	ResNet50	75.5	74.6	64.6
SurgFormer[a] (Ours)	Online	ResNet50	83.1	77.9	67.9
SurgFormer[a] (Ours)	Offline	ResNet50	**83.8**	**79.3**	**70.0**

[a] Indicates the multi-task model.

5.2. Implementation details

Training detail. Our model is implemented using PyTorch and trained in a distributed manner across four NVIDIA RTX 3090 GPUs. Each GPU processes a batch size of 2, with each batch containing $T = 32$ frames sampled sequentially from the same video. The interval between adjacent video sequences is set to 16 frames. For the TMVP-SurgVideo dataset, all input frames are resized to 320 × 320 and undergo data augmentations including random horizontal flip, random rotation, and color jitter before feeding into the model. SurgFormer is trained for 30 epochs, with the learning rate reduced by a factor of 0.1 at the 20th epoch. The AdamW optimizer (Loshchilov and Hutter, 2018) is adopted for model optimization, with the initial learning rate of 2×10^{-5} for the image encoder and 2×10^{-4} for the rest parts. For the cholec80 dataset, following the setting of previous works (Jin et al., 2020, 2022), we down-sample the video from 25 fps to 1 fps and resize frames into 250 × 250. Data augmentations, such as 224 × 224 cropping, random mirroring, and color jittering, are applied to the input images during training. The other training setting is the same as the TMVP-SurgVideo dataset. The loss weights for different task are set as $\lambda_1 = 1$, $\lambda_2 = 5$, $\lambda_3 = 1$, and $\lambda_4 = 5$. Since the loss curves exhibited stable convergence with consistent plateauing in later epochs in the training processes of both TMVP-SurgVideo and cholec80 datasets, we chose to use the model of the last epoch for performance validation and fine-tune the hyperparameters, such as learning rate and loss weights, based on the results on the validation set.

Inference detail. During inference, each video is divided into clips, with each clip containing 64 frames. Before model input, the images within each clip are resized to the same dimensions used during training. After processing all clips of a video, the prediction results are concatenated in order as the final predictions for the entire video. For the memory update in online inference, the probability threshold of $\tau = 0.7$ is employed.

5.3. Comparison with state-of-the-art methods

5.3.1. TMVP-SurgVideo dataset

In this part, we evaluate the performance of several representative SWA methods and our proposed SurgFormer on the four tasks of the TMVP-SurgVideo dataset. Additionally, we also evaluate the large pre-trained visual model DINOv2 (Oquab et al., 2024) on TMVP-SurgVideo. DINOv2 is pre-trained on 142M images and its backbone (ViT-B) is utilized to extract features from video frames, which are then used to produce the task predictions.

Phase Recognition. Since the phase recognition task is the mainstream task in the SWA field, more approaches are assessed and compared in this task. As reported in Table 2, SAHC achieves 75.5%, 74.6%, and 64.6% for PR, RE, and JA, respectively. The multi-task framework Trans-SVNet gets only 74.0% precision and 71.4% recall. With the same backbone, our online SurgFormer attains 83.1%, 77.9% and 67.9% for PR, ER, and JA, respectively, significantly outperforming previous methods and the DINOv2 model. Furthermore, our offline SurgFormer

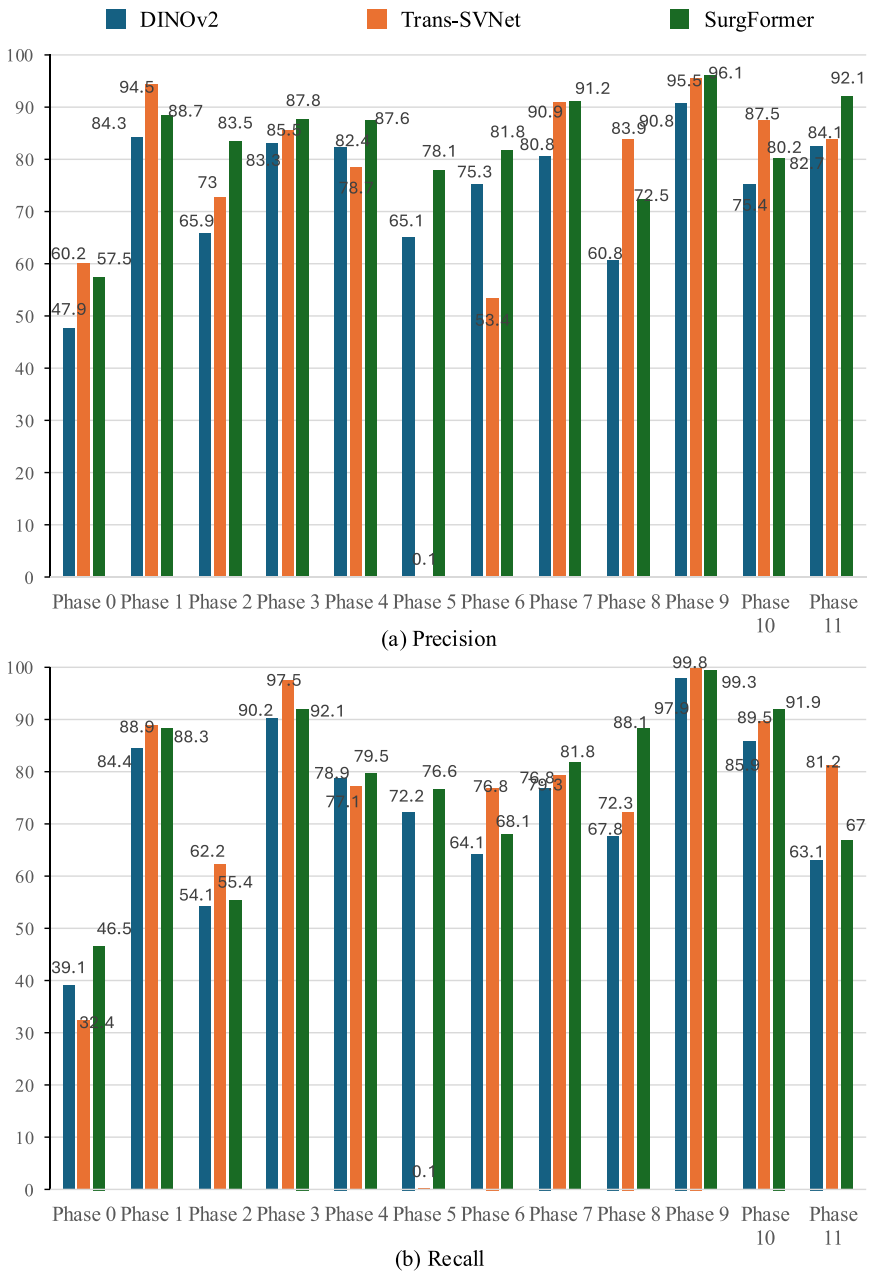

Fig. 8. The phase-wise results of (a) precision and (b) recall of three representative methods on TMVP-SurgVideo validation set.

further improves the performance across all metrics and achieves state-of-the-art results. Fig. 8 compares the precision and recall of each phase for DINOv2, Trans-SVNet, and online SurgFormer. It is evident that SurgFormer achieves precision exceeding 80% for most phases, with some even surpassing 90%. While Trans-SVNet exhibits similar precisions to SurgFormer on most phases, its precision on phase 5 is near zero, significantly impacting its overall precision. Fig. 9 presents the confusion matrix of the classification results, and we can see that the most easily misclassified phase is phase 0, which may be attributed to the fact that it lacks discriminative phase characteristics as a transition phase and appears randomly between phases.

Instrument Recognition. For the instrument recognition task, we compare our proposed SurgFormer with the multi-task Framework MTRCNet (Jin et al., 2020) and the DINOv2 model. As shown in Table 3, instruments that are frequently used or prominently visible, such as the surgical needle holder, endotherm knife, and thoracoscopic needle holder, have higher AP values across all methods. Both online and offline versions of SurgFormer outperform MTRCNet and the DINOv2 model in recognizing most instruments, with offline SurgFormer achieving the highest mAP. Notably, the DINOv2 model, despite lacking temporal information, still performs well in instrument recognition, suggesting that recognizing a single frame is crucial for this task. Fig. 10 depicts the confusion matrix of the classification results of each instrument, and it can be observed that the knife is the instrument most likely to be misclassified, corresponding to the results in Table 3.

Phase Anticipation. For the phase anticipation task, we evaluate DINOv2, Trans-SVNet, IIA-Net, and our SurgFormer on the TMVP-SurgVideo dataset, with results included in Table 4. It can be seen that both online and offline SurgFormer outperform the comparison methods across all three metrics, and the offline one achieves the best result.

Instrument Anticipation. The performance of DINOv2, IIA-Net, and our SurgFormer on the instrument anticipation task of the TMVP-SurgVideo dataset is evaluated, with the results presented in Table 4.

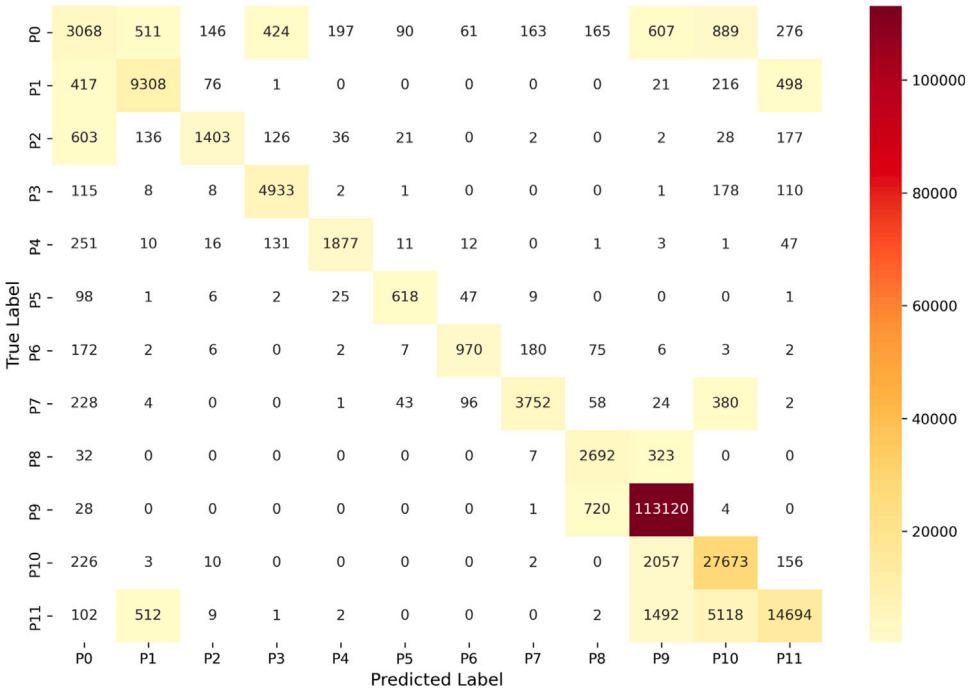

Fig. 9. The confusion matrix of phase recognition on TMVP-SurgVideo validation set.

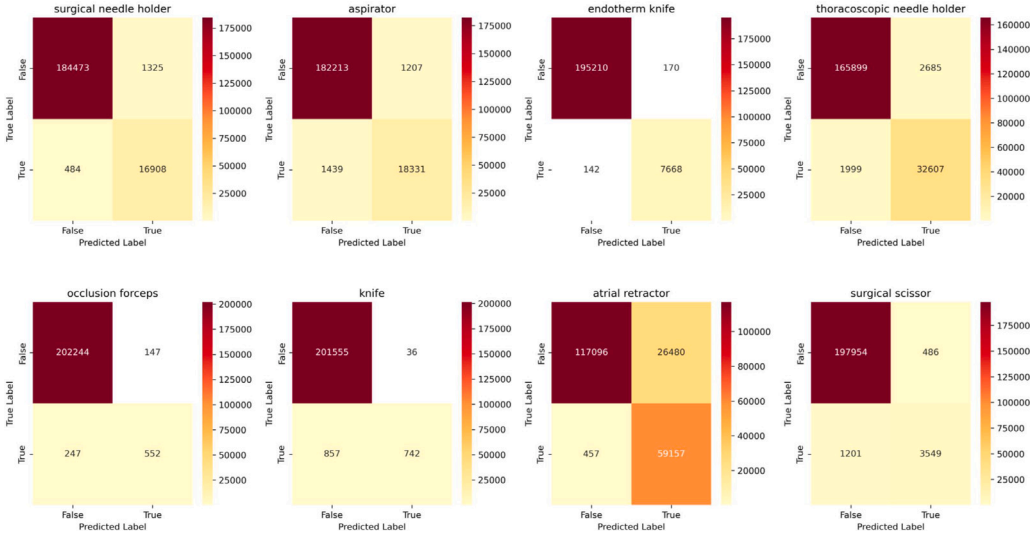

Fig. 10. The confusion matrix of instrument recognition on TMVP-SurgVideo validation set.

This task proves to be more challenging than phase anticipation. Both online and offline SurgFormer outperform IIA-Net and DINOv2 at all metrics, with the offline SurgFormer exhibiting a slight improvement over its online counterpart.

Fig. 11 represents partial results of phase and instrument anticipation on a validation video from TMVP-SurgVideo. Considering the readability of the visualization, we compare only the ground truth, and the results of our method as well as the current state-of-the-art method in each specific task. Compared to the previous method, the proposed SurgFormer exhibits smoother predictions that more closely align with the ground truth values for both phase and instrument anticipation.

Overall, a comprehensive analysis of the performance of existing methods on the four tasks of the TMVP-SurgVideo dataset reveals its challenging nature within the SWA field, warranting further investigation. Our proposed multi-task SurgFormer consistently outperforms state-of-the-art methods on all four tasks, demonstrating the effectiveness of multi-task collaboration within our query-based Transformer architecture. Furthermore, the offline SurgFormer, leveraging global information, consistently performs better than its online counterpart.

5.3.2. *External validation of TMVP*

In this part, we evaluate the performance of our proposed SurgFormer against the comparison methods across all four tasks of the external validation set. Table 5, Tables 6 and 7 present the results of phase recognition, instrument recognition and anticipation of both phase and instrument, respectively. It can be seen that SurgFormer achieves similar performance on the four tasks of the external validation set as it does on the TMVP-SurgVideo validation set, and all of them maintain a performance advantage over the comparison

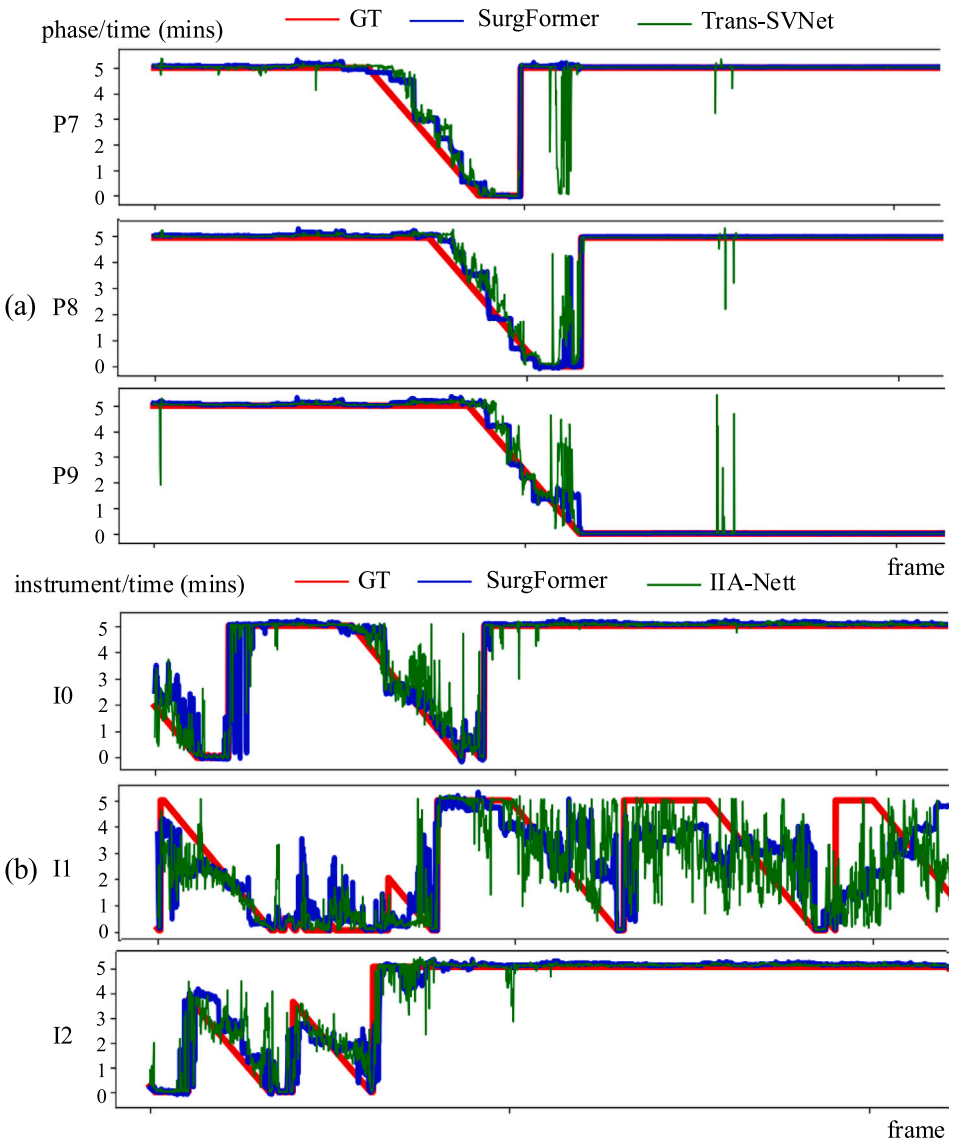

Fig. 11. Examples of (a) phase anticipation results and (b) instrument anticipation results of a validation video in TMVP-SurgVideo.

method. These results collectively demonstrate the model's strong generalizability in analyzing surgical videos from previously unseen surgeons.

5.3.3. Cholec80 dataset

We evaluate our proposed SurgFormer on the widely used Cholec80 benchmark and compare it with state-of-the-art SWA methods on the same four tasks. Particularly, for the anticipation task, we all set the horizon threshold h=5 min.

Phase Recognition. The performance of our two-mode SurgFormer on the phase recognition task of the Cholec80 test set is evaluated and compared with state-of-the-art approaches in Table 8. It can be observed that most of the methods achieve approximately 84% phase-level precision and higher recall, suggesting that the surgical scenes in the dataset are relatively straightforward. Our online and offline versions of SurgFormer outperform most representative SWA methods and achieve competitive performance with the elaborate SKiT model, which utilizes a powerful Vit-B backbone.

Instrument Recognition. Table 9 presents the instrument recognition performance of our models compared with existing approaches. Our offline SurgFormer gets the highest mAP, and both our SurgFormer and the multi-task framework MTRCNet could realize the mAP close to 90%. This high performance may be attributed to the relatively simple scenes of instrumentation usage in the Cholec80 dataset.

Phase Anticipation. The comparison of our SurgFormer and representative methods on this task is presented in Table 10. Both online and offline SurgFormer significantly outperforms previous multi-task models across all the MAE variants and the offline model delivers the best performance.

Instrument Anticipation. For the instrument anticipation task, we report the comparison results of our model and the multi-task framework IIA-Net in Table 10. It can be found that both online and offline versions of SurgFormer surpass IIA-Net in all three MAE metrics. We believe this advantage stems from the collaboration of more task information and effective temporal correlation learning.

5.4. Model analysis

We conduct extensive ablation experiments on the TMVP-SurgVideo dataset to validate the effectiveness of different key components in the proposed SurgFormer. For better clarity, we present the ablation results of the prevailing online mode SurgFormer. Phase, Ins, Phase_anti, and Ins_anti indicate phase recognition, instrument recognition, phase anticipation, and instrument anticipation, respectively.

Table 3
Comparison of average precision for instrument recognition on TMVP-SurgVideo validation set.

Instrument	DINOv2[a,b] (Oquab et al., 2024)	MTRCNet[a] (Jin et al., 2020)	SurgFormer[a]	SurgFormer[a,b]
Surgical needle holder	90.6	90.9	90.4	91.1
Thoracoscopic needle holder	85.8	87.1	88.1	88.5
Endotherm knife	95.9	95.6	96.1	96.0
Occlusion forceps	54.4	45.8	54.7	55.2
Aspirator	86.0	87.4	87.7	88.1
Knife	37.6	48.1	44.7	45.4
Atrial retractor	69.2	68.6	68.8	70.0
Surgical scissor	56.6	62.7	66.3	66.9
mAP	72.0	73.3	74.6	**75.2**

[a] Indicates the multi-task model.
[b] Indicates offline mode.

Table 4
Comparison of phase and instrument anticipations on TMVP-SurgVideo validation set.

Task	Method	inMAE ↓	eMAE ↓	pMAE ↓
Phase anticipation	IIA-Net[a] (Yuan et al., 2022)	1.13	1.22	1.65
	DINOv2[a,b] (Oquab et al., 2024)	1.07	1.24	1.40
	Trans-SVNet[a] (Jin et al., 2022)	1.06	1.24	1.41
	SurgFormer[a] (Ours)	0.95	0.92	1.22
	SurgFormer[a,b] (Ours)	**0.91**	**0.76**	**1.20**
Instrument anticipation	DINOv2[a,b] (Oquab et al., 2024)	1.41	1.72	1.45
	IIA-Net[a] (Yuan et al., 2022)	1.37	1.61	1.43
	SurgFormer[a] (Ours)	1.23	1.31	1.38
	SurgFormer[a,b] (Ours)	**1.19**	**1.22**	**1.36**

[a] Means the multi-task model.
[b] Indicates offline mode.

Table 5
Comparison results of phase recognition on the external validation set of TMVP-SurgVideo.

Method	Mode	Backbone	PR	RE	JA
SV-RCNet (Jin et al., 2018)	Online	ResNet50	70.7	67.9	58.5
TeCNO (Czempiel et al., 2020)	Online	ResNet50	72.9	72.2	60.6
TMRNet (Jin et al., 2021)	Online	ResNet50	73.1	71.6	60.1
Trans-SVNet[a] (Jin et al., 2022)	Online	ResNet50	73.2	70.9	60.5
MTRCNet[a] (Jin et al., 2020)	Online	ResNet50	74.2	72.2	61.6
DINOv2[a] (Oquab et al., 2024)	Offline	ViT-B	74.3	72.4	62.0
SAHC (Ding and Li, 2022)	Online	ResNet50	74.7	73.8	64.0
SurgFormer[a] (Ours)	Online	ResNet50	82.2	77.7	67.4
SurgFormer[a] (Ours)	Offline	ResNet50	**82.7**	**78.5**	**68.2**

[a] Indicates the multi-task model.

Table 6
Comparison of average precision for instrument recognition on the external validation set of TMVP-SurgVideo.

Instrument	DINOv2[a,b] (Oquab et al., 2024)	MTRCNet[a] (Jin et al., 2020)	SurgFormer[a]	SurgFormer[a,b]
Surgical needle holder	90.1	90.2	90.6	90.8
Thoracoscopic needle holder	85.4	86.3	87.8	88.1
Endotherm knife	95.2	94.6	95.6	95.5
Occlusion forceps	54.1	45.3	54.8	55.4
Aspirator	85.4	86.8	87.3	87.8
Knife	37.0	47.7	45.0	45.5
Atrial retractor	68.8	67.9	68.6	69.6
Surgical scissor	56.0	62.0	65.5	66.5
mAP	71.5	72.6	74.4	**74.9**

[a] Indicates the multi-task model.
[b] Indicates offline mode.

Table 7
Comparison of phase and instrument anticipations on the external validation set of TMVP-SurgVideo.

Task	Method	inMAE ↓	eMAE ↓	pMAE ↓
Phase anticipation	IIA-Net[a] (Yuan et al., 2022)	1.16	1.24	1.67
	DINOv2[a,b] (Oquab et al., 2024)	1.12	1.22	1.41
	Trans-SVNet[a] (Jin et al., 2022)	1.10	1.25	1.43
	SurgFormer[a] (Ours)	1.01	0.97	1.23
	SurgFormer[a,b] (Ours)	**0.97**	**0.86**	**1.18**
Instrument anticipation	DINOv2[a,b] (Oquab et al., 2024)	1.39	1.74	1.44
	IIA-Net[a] (Yuan et al., 2022)	1.33	1.63	1.42
	SurgFormer[a] (Ours)	1.24	1.29	1.36
	SurgFormer[a,b] (Ours)	**1.18**	**1.24**	**1.33**

[a] Means the multi-task model.
[b] Means offline mode.

Table 8
Comparison results of phase recognition on Cholec80 test set.

Method	Mode	Backbone	PR	RE	JA
TeCNO (Czempiel et al., 2020)	Online	ResNet50	80.3	82.5	68.4
MTRCNet[a] (Jin et al., 2020)	Online	ResNet50	81.2	82.7	71.4
TMRNet (Jin et al., 2021)	Online	ResNet50	84.1	84.2	72.1
Trans-SVNet[a] (Jin et al., 2022)	Online	ResNet50	84.7	83.6	72.5
SAHC (Ding and Li, 2022)	Online	ResNet50	84.3	86.1	74.6
SKiT (Liu et al., 2023)	Online	ViT-B	84.6	**88.5**	**76.7**
SurgFormer[a]	Online	ResNet50	84.7	87.8	75.8
SurgFormer[a]	Offline	ResNet50	**85.1**	88.3	76.3

[a] Indicates the multi-task model.

5.4.1. Information interaction module

Here, we investigate the impact of the two layers in the information interaction module on model performance. As shown in Table 11, removing the information interaction module from SurgFormer leads to a significant drop in performance across all four metrics. Adding either the intra-frame TLII layer or the inter-frame TCL layer independently to the model results in a substantial performance increase, with the inter-frame TCL layer contributing a greater improvement than the intra-frame TLII layer. This suggests that learning temporal correlations within the information interaction module is slightly more impactful

Table 9
Comparison of AP for instrument recognition on Cholec80 test set.

Instrument	EndoNet[b] (Twinanda et al., 2017)	MTRCNet[a] (Jin et al., 2020)	SurgFormer[a]	SurgFormer[ab]
Grasper	84.8	84.7	85.6	86.0
Bipolar	86.9	90.1	90.0	90.2
Hook	95.6	95.6	97.7	97.8
Scissors	58.6	86.7	86.8	87.1
Clipper	80.1	89.8	90.6	90.9
Irrigator	74.4	88.2	89.8	90.0
Specimen bag	86.8	85.1	87.0	87.2
mAP	81.0	89.1	89.7	**89.9**

[a] Means the multi-task model.
[b] Means offline mode.

Table 10
Comparison of phase and instrument anticipations on Cholec80 test set.

Task	Method	inMAE ↓	eMAE ↓	pMAE ↓
Phase anticipation	IIA-Net[a] (Yuan et al., 2022)	1.28	1.60	1.78
	Trans-SVNet[a] (Jin et al., 2022)	1.25	1.53	1.72
	SurgFormer[a] (Ours)	1.04	1.07	1.47
	SurgFormer[ab] (Ours)	**0.99**	**1.01**	**1.32**
Instrument anticipation	IIA-Net[a] (Yuan et al., 2022)	1.56	2.23	1.85
	SurgFormer[a] (Ours)	1.28	1.31	1.35
	SurgFormer[ab] (Ours)	**1.22**	**1.28**	**1.32**

[a] Indicates the multi-task model.
[b] Means offline mode.

Table 11
Ablation study on the information interaction module (IIM). Intra refers to the intra-frame TLII layer. Inter represents the inter-frame TCL layer.

IIM		Phase	Ins	Phase_ant	Ins_ant
Intra	Inter	PR	mAP	inMAE↓	inMAE ↓
×	×	77.1	71.9	1.15	1.42
✓	×	79.6	72.4	1.08	1.35
×	✓	81.4	73.3	1.01	1.30
✓	✓	**83.1**	**74.6**	**0.95**	**1.23**

Table 12
Ablation study on the multi-task collaboration mechanism.

Task				Phase	Ins	Phase_anti	Ins_anti
Phase	Ins	Phase_anti	Ins_anti	PR	mAP	inMAE ↓	inMAE ↓
✓				60.6	–	–	–
	✓			–	73.0	–	–
		✓		–	–	1.12	–
			✓	–	–	–	1.33
✓	✓			80.2	74.0	–	–
✓		✓		79.6	–	1.01	–
✓			✓	76.7	–	–	1.29
	✓	✓		–	73.2	1.08	–
	✓		✓	–	73.7	–	1.24
		✓	✓	–	–	1.05	1.26
✓	✓	✓	✓	**83.1**	**74.6**	**0.95**	**1.23**

than facilitating information interaction between tasks in our multi-task architecture. SurgFormer achieves the best performance when both layers are included, demonstrating the effectiveness of the module.

5.4.2. Multi-task collaboration

We explore the performance of SurgFormer with different task combinations on the TMVP-SurgVideo dataset, including single-task, two-task, and four-task. As shown in Table 12, the performance of single-task models is inferior to that of multi-task models, particularly in the phase recognition task. Furthermore, among the two-task models, the combination of phase and instrument recognition achieves better performance in both tasks. Additionally, phase recognition information improves phase anticipation and instrument recognition improves instrument anticipation. Finally, the experimental results show that the SurgFormer with four-task collaboration achieves the best performance across all four tasks.

Table 13
Ablation study on the probability threshold of memory update during online inference.

Probability threshold	Phase	Ins	Phase_ant	Ins_ant
	PR	mAP	inMAE↓	inMAE↓
None	82.2	74.1	1.08	1.31
0.5	82.8	74.4	1.00	1.25
0.6	82.9	74.4	0.98	1.25
0.7	**83.1**	**74.6**	**0.95**	**1.23**
0.8	83.1	74.5	0.96	1.23
0.9	82.8	74.3	0.98	1.26

Table 14
Ablation study on the unit number of information interaction module.

Unit number	Phase	Ins	Phase_ant	Ins_ant	FPS
	PR	mAP	inMAE↓	inMAE↓	
1	82.6	73.8	1.05	1.30	383.2
2	**83.1**	**74.6**	**0.95**	1.23	376.5
3	83.0	74.3	0.97	**1.20**	367.8

5.4.3. Memory update threshold

Here, we investigate the impact of the probability threshold value of memory update during online inference on the model's performance. As shown in Table 13, the "None" indicates that we do not assess the phase recognition probability and instead directly store the task embeddings into the memory bank. Our analysis reveals that SurgFormer achieves optimal performance when the threshold is set to 0.7. In contrast, performance slightly degrades with other threshold values, while the degradation is more pronounced when the threshold is set to None. These findings underscore the importance of selecting high-quality task embeddings during memory updates for online inference.

5.4.4. Number of units in information interaction module

In this study, we investigate the influence of the number of units within the information interaction module on the performance of SurgFormer. As represented in Table 14, we evaluate the models with 1, 2, and 3 units, corresponding to 2, 3, and 4 decoder layers in the Transformer. The experimental results show that SurgFormer with 1 unit exhibits inferior performance on the four tasks, while the models with 2 and 3 units achieve improved performance. Considering the balance between performance and running efficiency, we chose 2 units for the information interaction module.

5.4.5. Skill level of the surgeon

In this part, we explore the performance difference of our model on the videos of expert and novice surgeons of the TMVP-SurgVideo validation set across the four tasks. The quantitative comparison results are shown in Table 15. The results demonstrate that while our model achieves marginally better performance on expert surgeons' videos, it maintains comparable effectiveness when analyzing novice surgeons' procedures. Notably, the performance gap between these two groups remains relatively small, indicating robust generalization across skill levels.

Table 15
Comparison of the expert and novice surgeons of the four tasks on TMVP-SurgVideo validation set.

Phase recognition					Instrument Recognition				
Surgeon	Method	PR	RE	JA	Surgeon	Method	mAP		
Expert	SurgFormer[a]	83.3	78.0	68.2	Expert	SurgFormer[a]	74.8		
	SurgFormer[b]	**84.0**	**79.5**	**70.2**		SurgFormer[b]	**75.4**		
Novice	SurgFormer[a]	82.7	77.4	67.5	Novice	SurgFormer[a]	74.2		
	SurgFormer[b]	83.5	79.1	69.8		SurgFormer[b]	74.9		
Phase anticipation					Instrument anticipation				
Surgeon	Method	inMAE	eMAE	pMAE	Surgeon	Method	inMAE	eMAE	pMAE
Expert	SurgFormer[a]	0.92	0.90	1.18	Expert	SurgFormer[a]	1.20	1.28	1.36
	SurgFormer[b]	**0.89**	**0.75**	**1.14**		SurgFormer[b]	**1.16**	**1.18**	**1.33**
Novice	SurgFormer[a]	0.97	0.95	1.25	Novice	SurgFormer[a]	1.25	1.33	1.42
	SurgFormer[b]	0.92	0.78	1.22		SurgFormer[b]	1.23	1.25	1.40

[a] Means online model.
[b] Means offline mode.

6. Discussion

Our work introduces the first large-scale dataset TMVP-SurgVideo and a versatile multi-task method for SWA on thoracoscopic mitral valvuloplasty. In the following part, we will discuss the achievements and limitations of the dataset and the proposed SurgFormer, and some future directions.

6.1. Dataset

One achievement of the dataset is having a large number of recorded surgical videos of complex TMVP with variability. Compared to laparoscopic surgeries, thoracoscopic cardiac surgery presents greater technical complexity and operational challenges. Laparoscopic procedures typically involve working with static abdominal organs in a relatively spacious operative field, often focusing on tissue mobilization and lesion excision. In contrast, thoracoscopic cardiac surgery requires operating in a confined and dynamic environment. This is particularly evident in mitral valve surgery, where surgeons must perform not only tissue mobilization and pathological excision but also high-precision suturing and reconstructive procedures on delicate, demonstrating greater operational complexity. Moreover, these surgeries were performed by four surgeons with different surgical operating habits and proficiency, and the various conditions that may be encountered during surgery, such as bleeding, artificial valves suture failures, etc., can lead to varying surgical durations, e.g., the shortest video duration in our dataset is 43.78 min, and the longest one reaches 300.92 min. Besides, the angle and light intensity at which the camera captured the surgical procedure also show variability. This emphasizes the necessity of training multi-task models on datasets with variability to improve performance and generalizability.

Despite these accomplishments, there are also some limitations to the dataset. First, due to the lack of a corresponding gold standard as a reference, the division of phases in the dataset is based on the experience of three expert surgeons, thus it may be divergent for other surgeons. Second, for the annotation of instruments, due to the occlusion between instruments and the limitation of the camera view, sometimes only a small portion of the instruments are visible, and then it is difficult to judge whether these instruments are present or not. Third, during the construction of the dataset, due to the long duration of each video, we reduced the frame rate of the videos to 1FPS in order to save storage space and improve the annotation efficiency, which may lead to the loss of information, thus leading to the performance degradation of some SWA methods.

6.2. Method

The multi-task framework can output more information about different aspects of the surgical procedure compared to the single-task framework, which is also conducive to a comprehensive understanding of the surgical process. Finding the intrinsic correlation between multiple tasks is the guiding principle in building a multi-task framework to address the various challenges in the dataset. As presented in Fig. 2, phases 10 and 11 exhibit similarities in visual background and instrument usage, potentially leading to misclassification. However, as depicted in rows 1 and 2 of Fig. 2, when we incorporate the instrument usage and its duration, i.e., recognition and anticipation, information into the model, phase discrimination will be enhanced, since the endotherm knife will not be used in phase 10 and the instrument usage shows continuity. Furthermore, the strong correlation between phase division and instrument usage implies that accurate phase recognition can, in turn, facilitate instrument recognition and anticipation. Similarly, phase 11 presents internal variance, as seen in rows 2 and 3 of Fig. 2. While if the duration information of a phase is introduced, it will impose constraints on the temporal continuity of phases, thereby enhancing the phase recognition performance, which in turn can improve the accuracy of phase anticipation. Besides, the temporal continuity of phases and instruments can help address some complex scene problems presented in Fig. 3.

In terms of architecture design, unlike previous approaches that lack explicit inter-task information interaction and predict multiple tasks directly from a single feature, our approach aims to decouple the feature representation of each task and achieve explicit inter-task information collaboration at less cost. This decoupled design and collaboration mechanism both reduce the difficulty of feature learning for specific tasks and ensure effective information interaction between multiple tasks. We believe this is the core reason for the better performance of our model, as we use the same image encoder (ResNet-50) as the previous methods. On the other hand, more powerful image encoders, such as ViT (Dosovitskiy et al., 2021) and Swin-Transformer (Liu et al., 2021), may further improve the performance.

Despite the high performance of our proposed SurgFormer, it still encounters some limitations. First, our model does not address the severe class imbalance that exists in the dataset. For example, as shown in Table 3, the average precision of knife and occlusion forceps is only 37.6% and 54.4%, respectively. We think this may be due to their low proportion of sample numbers in the entire dataset: the proportion of sample numbers of knife is 2827/429,494, and occlusion forceps is 2079/429,494. How to address the class imbalance issue may be a possible research direction in the future. Second, currently, the information interaction strategy in SurgFormer adopts the intra-frame task-level information interaction and inter-frame temporal correlation learning, while more information interaction strategies for task embedding have not been fully explored. For instance, information interaction between

different task embeddings at different frames and different information interaction mechanisms except the self-attention mechanism. Third, due to the constraints of hardware, the offline-mode model is unable to accept the whole long video as input at once during the inference process, so the long video needs to be split into fixed-length video clips to be input into the model, which may limit the model's ability to perceive the global contextual information and affect the performance.

Finally, we conclude with some directions for future work on the TMVP-SurgVideo and the proposed SurgFormer. First, based on the TMVP-SurgVideo dataset, more SWA tasks, such as fine-grained step recognition task and high-level surgical skill assessment task, could be further explored. However, this may require the corresponding annotation work to be performed with the help of some professional surgeons. Second, the proposed multi-task model could be applied and improved in other surgical scenes. Since each task is associated with a specific task embedding for uniform decoding, it is easy for SurgFormer to adapt to a two-task or three-task framework. Moreover, the lightweight SWA model could also be a future direction for edge-side model deployment.

7. Conclusion

This paper introduces TMVP-SurgVideo, the first large-scale TMVP dataset with annotations of four key tasks related to the surgical phase and instrument. This dataset expands the scope of SWA research to include complex cardiac surgery, offering a valuable resource for intelligent surgery. To provide a comprehensive SWA system for TMVP-SurgVideo, we propose SurgFormer, a pioneering query-based Transformer framework that can simultaneously realize four SWA tasks. SurgFormer learns a low-dimensional task embedding for the prediction of each task and incorporates an information interaction module in the decoder that effectively collaborates task representations within each frame and captures temporal correlations across frames. The flexible architecture of SurgFormer allows for both offline and online inference modes, making it adaptable to various SWA applications. Experimental results on the TMVP-SurgVideo and Cholec80 datasets demonstrate the challenging nature of TMVP-SurgVideo and highlight the effectiveness of SurgFormer.

CRediT authorship contribution statement

Meng Lan: Writing – original draft, Visualization, Methodology. **Weixin Si:** Writing – review & editing, Methodology, Formal analysis. **Xinjian Yan:** Validation, Investigation, Data curation. **Xiaomeng Li:** Resources, Writing – review & editing, Project administration, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by grants from the National Natural Science Foundation of China under Grant No. 62306254, grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: T45-401/22-N).

References

Blum, T., Feußner, H., Navab, N., 2010. Modeling and segmentation of surgical workflow from laparoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 400–407.

Cao, J., Yip, H.-C., Chen, Y., Scheppach, M., Luo, X., Yang, H., Cheng, M.K., Long, Y., Jin, Y., Chiu, P.W.-Y., et al., 2023. Intelligent surgical workflow recognition for endoscopic submucosal dissection with real-time animal study. Nat. Commun. 14 (1), 6676.

Chen, Y., He, S., Jin, Y., Qin, J., 2023. Surgical activity triplet recognition via triplet disentanglement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 451–461.

Committee, C.M.I.C.S., 2023. Endoscopic mitral valve replacement & repair-benchmarking CMICS 2022. Chin. J. Thorac. Cardiovasc. Surg. 39 (5), 257–264.

Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N., 2020. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 343–352.

Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N., 2021. Opera: Attention-regularized transformers for surgical phase recognition. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 604–614.

Ding, X., Li, X., 2022. Exploring segment-level semantics for online phase recognition from surgical videos. IEEE Trans. Med. Imaging 41 (11), 3309–3319.

Ding, X., Yan, X., Wang, Z., Zhao, W., Zhuang, J., Xu, X., Li, X., 2023. Less is more: Surgical phase recognition from timestamp supervision. IEEE Trans. Med. Imaging 42 (6), 1897–1910.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Farha, Y.A., Gall, J., 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3575–3584.

Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.-A., 2021. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 593–603.

Gibaud, B., Forestier, G., Feldmann, C., Ferrigno, G., Gonçalves, P., Haidegger, T., Julliard, C., Katić, D., Kenngott, H., Maier-Hein, L., et al., 2018. Toward a standard ontology of surgical process models. Int. J. Comput. Assist. Radiol. Surg. 13, 1397–1408.

Graves, A., Graves, A., 2012. Long short-term memory. Supervised Seq. Label. Recurr. Neural Netw. 37–45.

Hayashi, Y., Nakamura, Y., Hirano, T., Ito, Y., Watanabe, T., 2021. Cumulative sum analysis for the learning curve of minimally invasive mitral valve repair. Heart Vessels 36, 1584–1590.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.-W., Heng, P.-A., 2018. SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans. Med. Imaging 37 (5), 1114–1126.

Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.-W., Heng, P.-A., 2020. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med. Image Anal. 59, 101572.

Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.-A., 2021. Temporal memory relation network for workflow recognition from surgical video. IEEE Trans. Med. Imaging 40 (7), 1911–1923.

Jin, Y., Long, Y., Gao, X., Stoyanov, D., Dou, Q., Heng, P.-A., 2022. Trans-svnet: hybrid embedding aggregation transformer for surgical workflow analysis. Int. J. Comput. Assist. Radiol. Surg. 17 (12), 2193–2202.

Lalys, F., Jannin, P., 2014. Surgical process modelling: a review. Int. J. Comput. Assist. Radiol. Surg. 9, 495–511.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Liu, Y., Huo, J., Peng, J., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S., 2023. Skit: a fast key information video transformer for online surgical phase recognition. In: Proceedings of the International Conference on Computer Vision. pp. 21074–21084.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Loshchilov, I., Hutter, F., 2018. Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations. ICLR.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al., 2022a. Surgical data science–from concepts toward clinical translation. Med. Image Anal. 76, 102306.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al., 2022b. Surgical data science–from concepts toward clinical translation. Med. Image Anal. 76, 102306.

Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al., 2017. Surgical data science for next-generation interventions. Nat. Biomed. Eng. 1 (9), 691–696.

Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., et al., 2021. Heidelberg colorectal data set for surgical data science in the sensor operating room. Sci. Data 8 (1), 101.

Nwoye, C.I., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N., 2020. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 364–374.

Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N., 2022. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Med. Image Anal. 78, 102433.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al., 2024. DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. 1–31.

Otto, C.M., Nishimura, R.A., Bonow, R.O., Carabello, B.A., Erwin III, J.P., Gentile, F., Jneid, H., Krieger, E.V., Mack, M., McLeod, C., et al., 2021. 2020 ACC/AHA guideline for the management of patients with valvular heart disease: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J. Am. Coll. Cardiol. 77 (4), 450–500.

Quellec, G., Lamard, M., Cochener, B., Cazuguel, G., 2014. Real-time segmentation and recognition of surgical tasks in cataract surgery videos. IEEE Trans. Med. Imaging 33 (12), 2352–2360.

Ramesh, S., Dall'Alba, D., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Fiorini, P., Padoy, N., 2021. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. Int. J. Comput. Assist. Radiol. Surg. 16, 1111–1119.

Rivoir, D., Bodenstedt, S., Funke, I., von Bechtolsheim, F., Distler, M., Weitz, J., Speidel, S., 2020. Rethinking anticipation tasks: Uncertainty-aware anticipation of sparse surgical instrument usage for context-aware assistance. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 752–762.

Shah, N.A., Sikder, S., Vedula, S.S., Patel, V.M., 2023. GLSFormer: Gated-long, short sequence transformer for step recognition in surgical videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 386–396.

Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N., 2023. Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. Int. J. Comput. Assist. Radiol. Surg. 18 (6), 1053–1059.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N., 2017. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging 36, 86–97.

Valderrama, N., Ruiz Puentes, P., Hernández, I., Ayobi, N., Verlyck, M., Santander, J., Caicedo, J., Fernández, N., Arbeláez, P., 2022. Towards holistic surgical scene understanding. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 442–452.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the International Conference on Neural Information Processing Systems. pp. 5998–6008.

Wagner, M., Müller-Stich, B.-P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al., 2023. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Med. Image Anal. 86, 102770.

Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.-H., Dou, Q., Liu, Y., 2022. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 486–496.

Yi, F., Jiang, T., 2019. Hard frame detection and online mapping for surgical phase recognition. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 449–457.

Yuan, K., Holden, M., Gao, S., Lee, W., 2022. Anticipation for surgical workflow through instrument interaction and recognized signals. Med. Image Anal. 82, 102611.

Yue, W., Liao, H., Xia, Y., Lam, V., Luo, J., Wang, Z., 2023. Cascade multi-level transformer network for surgical workflow analysis. IEEE Trans. Med. Imaging 42 (10), 2817–2831.

Zhong, L., Huang, H., 2023. Application and development of totally thoracoscopic mitral valve plasty. Chin. J. Clin. Thorac. Cardiovasc. Surg. 458–463.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2021. Deformable DETR: Deformable transformers for End-to-End Object Detection. In: Proceedings of the International Conference on Learning Representations.