This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SKiT: a Fast Key Information Video Transformer for Online Surgical Phase Recognition

Yang Liu, Jiayu Huo, Jingjing Peng, Rachel Sparks,
Prokar Dasgupta, Alejandro Granados, Sebastien Ourselin*
King's College London
{yang.9.liu, firstname.lastname}@kcl.ac.uk

Abstract

This paper introduces SKiT, a fast Key information Transformer for phase recognition of videos. Unlike previous methods that rely on complex models to capture long-term temporal information, SKiT accurately recognizes high-level stages of videos using an efficient ***key pooling*** *operation. This operation records important key information by retaining the maximum value recorded from the beginning up to the current video frame, with a time complexity of $\mathcal{O}(1)$. Experimental results on Cholec80 and AutoLaparo surgical datasets demonstrate the ability of our model to recognize phases in an online manner. SKiT achieves higher performance than state-of-the-art methods with an accuracy of 92.5% and 82.9% on Cholec80 and AutoLaparo, respectively, while running the temporal model* ***eight*** *times faster (7ms v.s. 55ms) than LoViT, which uses ProbSparse to capture global information. We highlight that the inference time of SKiT is constant, and independent from the input length, making it a stable choice for keeping a record of important global information, that appears on long surgical videos, essential for phase recognition. To sum up, we propose an effective and efficient model for surgical phase recognition that leverages key global information. This has an intrinsic value when performing this task in an online manner on long surgical videos for stable real-time surgical recognition systems.*

1. Introduction

Surgical Artificial Intelligence uses data to understand surgical workflows, evaluate surgeon performance, and provide assistance to surgeons in real time[30]. One of the core tasks towards achieving these aims is the recognition of the transitions of high-level stages of surgery, a problem coined Surgical Phase Recognition [15]. Accurate predictions of what surgical phase a part of a video relates to might benefit the provision of automated and improved feedback for trainees [12, 24], the potential to optimize surgical workflows [32], and the retrospective review of a particular phase from surgical video. While past and future information is used for phase recognition of a particular video frame in an offline manner, only past information is used for classifying in an online manner the phase where the last (current) video frame locates. Although online recognition is more challenging, it could help alert surgeons [33] and support decision-making [8] in real-time during surgery. This paper focuses on online phase recognition.

Early work in surgical phase recognition proposed workflow recovery models using Dynamic Time Warping with temporal registration [1], graphical probabilistic models based on Hidden Markov Models (HMM) [2, 3], rule-based interpretation models for context-awareness using ontologies [23], and machine learning models for phase recognition using Support Vector Machines and Random Forests [15]. Although these methods are mathematically rigorous, the use of hand-crafted features is specific to the surgery type and leads to a design that is not fully generalisable. Deep learning brought in new methods for surgical phase recognition, which allows for more sophisticated spatio-temporal feature extraction mechanisms. While additional information, such as surgical tools presence, is considered by other methods in a multi-task learning manner to improve accuracy [21, 37], annotation requirements limit their influence on online recognition.

Online surgical phase recognition requires models that can capture long-range temporal dependencies since the duration of surgical videos could range from 40 minutes to a few hours. Single-task surgical phase recognition models that add temporality can be broadly categorized into three types which use recurrent neural networks (RNN) [34], convolution neural networks (CNN) [27], or Transformers [39]. However, while RNNs, including Long Short-Term Memory (LSTM), struggle with modelling long-term dependencies due to their sequential nature, CNN-based methods such as Temporal Convolutional Networks (TCNs)

*Corresponding author

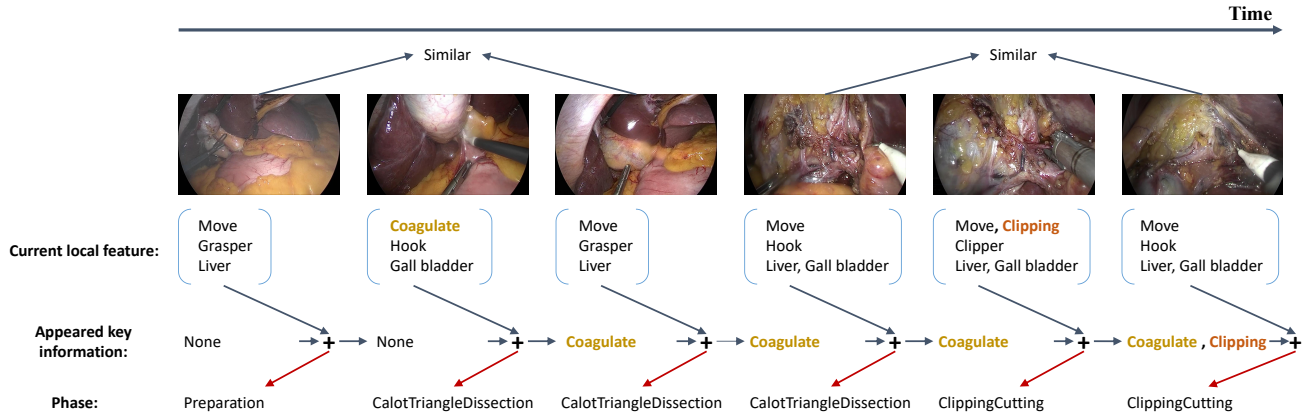


Figure 1: The *Key-recorder* aggregates previous and current local features to predict the current phase. The surgical phase is updated only when new key information is recognised. Working principle of our *Key-recorder* on an illustrative video stream to record the global appeared key information (2nd row) together with the current local feature (1st row) for recognising the phase (last row) of a current image frame.

use dilated convolution with fixed-sized filters to capture long temporal information, which can be problematic for long sequences. Moreover, dilated convolution can result in information loss while processing long sequences due to sparse sampling of input features. In contrast, Transformer-based methods have shown promising results as they can capture relationships between different tokens in a sequence, regardless of positions. This makes it easier to model long-term dependencies. One disadvantage is that time and memory complexity of the self-attention mechanism used by Transformers is quadratic, which limits their usefulness for long videos. Even with the use of *ProbSparse* attention [44] that decreases the complexity of self-attention, inference time is still related to the input length and could be time-consuming for long sequences. Furthermore, complex temporal models, especially for time-series data, may face greater challenges in retaining redundant information, which can cause overfitting. This is due to the frequent presence of autocorrelation and periodicity in time-series data, which may exacerbate overfitting when the model is overly complex. Further research is required to efficiently and effectively capture global information while ensuring processing time is independent of the input length.

To process the input with varying lengths while maintaining efficiency and effectiveness, it is important to consider previously captured information along with that appearing in the current frame. As depicted in Figure 1, certain local temporal features, such as 'Move grasper', may appear in multiple phases, such as 'Preparation' and 'CalotTriangleDissection', and the key information that distinguishes them is crucial. Assume that we recorded the appeared key information where it is from the beginning of the video to the current frame (not included), and we can efficiently combine it with the current local feature to update the appeared key information of the next frame, enabling us to reuse captured information, rather than searching it from the beginning of the video. Motivated by this, we propose a *fast **K**ey **i**nformation video **T**ransformer* (abbreviated as **SKiT**), which has the ability to record global appeared key information along the temporal dimension by an efficient and effective key pooling operation. After knowing the global appeared key information and current local fine-grained feature, SKiT could recognise the current phase accurately. Our main contributions are: 1) a new key pooling method that globally records important key events within $\mathcal{O}(1)$ time complexity, 2) a more efficient and stable approach that ensures inference time is not affected by video length, and 3) an efficient and accurate model that maintains state-of-the-art performance. The code website: https://github.com/MRUIL/SKiT.

2. Related Work

Multi-task learning methods. With the advent of deep learning, research has transformed into using only video as input, rather than operating on extra information that might be inconvenient to collect. Some works jointly learn phase recognition and tool presence detection through shared features [21]. Twinanda *et al*. [37] presented the multi-task network EndoNet which had two branches that shared early layers for visual feature extraction. However, it detached crucial temporal dependencies from the unified framework. In order to improve EndoNet's ability to understand tem-

poral context, Twinanda [35] replaced HMM with LSTM gates [19]. Zisimopoulos *et al*. [45] proposed a two-stage approach for cataract video analysis, using a ResNet [18] for tool presence recognition and an RNN for phase recognition, achieving promising results. Nakawala *et al*. [31] introduced a network integrating deep models with ontology and production rules to recognize surgical contexts. Jin *et al*. [21] utilized a correlation loss to enhance the performance of both tool presence detection and surgical phase recognition tasks by leveraging their correlation.

Single-task learning methods. Single-task models for phase recognition have been proposed since multi-task learning methods require extra tool annotations, which limits model training scenarios and increases annotation workload. Some methods [20, 41, 13] employed LSTMs [19] for temporal feature aggregation of surgical videos. However, as mentioned before, LSTMs suffer from the vanishing gradients problem, which limits their ability to capture long-term dependencies, particularly in surgical videos that can last for hours [5]. Although TMRNet [22] attempted to use a non-local operator to establish the relationship between the current feature and the global feature sequence, this approach was limited in addressing the issue because the global features affected the current feature independently rather than collaboratively. Czempiel *et al*. [5] presented TeCNO, based on TCNs [26, 11], that are able to capture long-term temporal correlations. However, by essentially adapting dilated convolutions [38] for long sequences using TCNs, the increased receptive field obtained through dilation can result in a loss of fine-grained relationships between more distant time steps. Moreover, the receptive field of TCNs is limited by the size of the convolutional filters, which can be problematic when dealing with long sequences. Transformer-based [39] models have also been proposed for general computer vision tasks, such as action recognition[10, 28] and action anticipation [16], which share similarities with surgical phase recognition. However, these methods have been designed for short video inputs. For surgical phase recognition, Trans-SVNet [14] aims to fuse spatial and temporal features by developing a small Transformer-based fusion head. However, the dilated convolution structure in TCNs still leads to fine-grained temporal information loss and fixed reception field. Czempiel *et al*. [6] also introduced a Transformer-based model for aggregating temporal features, but the quadratic time and memory complexity caused by its self-attention mechanism pose a particular problem for long surgical videos. Most recently, Long Video Transformer (LoViT) [29] employed $ProbSparse$ [44] to decrease the time complexity of the vanilla Transformer to capture global information. However, the inference time of these temporal models is highly dependent on the length of the input sequence because they require a global reception field that needs additional computations to process long-term features. As the length of surgical videos increases, their inference time also increases, which is undesirable for building stable real-time surgical recognition systems. A naive solution used by these models consists of fixing the input length and dropping early input frames, which causes preceding information to be lost. The TeSTra [43] proposed for nature video action recognition update global features by selectively discarding older information, achieving an efficient time complexity of $\mathcal{O}(1)$. However, it's worth noting that older long-term events retain significance in the context of current phase recognition.

Building on previous work, our research also leverages the Transformer network to extract local fine-grained temporal features and spatial information. However, to process the variable long surgical video and capture global temporal information, we needed an effective and efficient solution. To address this, we propose **key pooling**, which no longer needs to build a complex global temporal model. This idea was inspired by CornerNet [25], which proposed corner pooling to detect image objects as paired key points. The main idea behind corner pooling is to take the maximum value from different boundaries of the image to the current position and add them together. For online video streams, our key pooling approach records key information that appears along the time dimension from the beginning to the current frame. This allows us to reuse previous frame-wise output while capturing global key information up to the current frame.

3. Methods

Figure 2 illustrates our proposed SKiT architecture. Our model takes a video stream $X_t = \{x_i\}_{i=1}^{t}$ as input to recognise the surgical phase $\hat{p_t}$ of the current frame x_t in an online manner, where x_i is the i-th video frame contained in X_t. We propose an architecture with a novel *Key-recorder* that captures global information by recording key events, thereby eliminating the need for costly aggregation operations. This approach addresses the computational limitations that result from large-sized local and global feature aggregator operations inherent in LoViT [29], which begins with extracting spatial features, followed by the extraction of small- and large-sized local temporal features, and then aggregates global long sequence features using a global temporal feature aggregator while performing multi-scale fusion with local features. We describe our proposed architecture in the following sections.

3.1. Spatial Feature Extractor

Surgical videos can last for up to a few hours and exhibit strong dependence among different phases. Therefore, it is essential for a recognition model to have the ability to process long video inputs. However, training such a model in an end-to-end manner is challenging. Previous

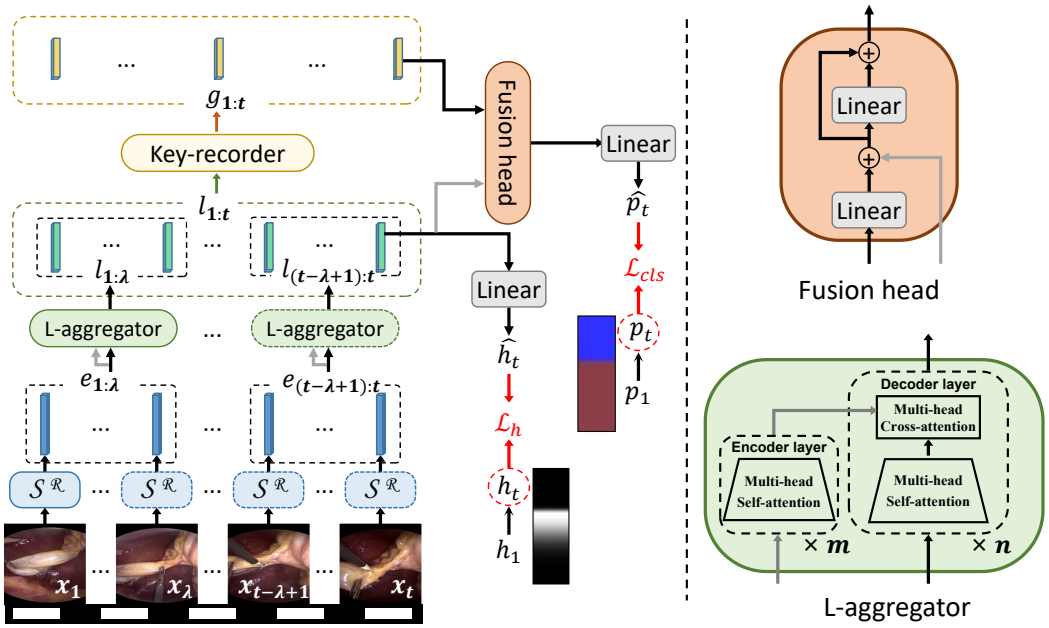


Figure 2: **SKiT architecture.** First, a spatial feature extractor $\mathcal{S}^{\mathcal{R}}$ is used to independently extract a sequence of spatial features $(e_1, ..., e_t)$ from a t-frames video stream $(x_1, ..., x_t)$. A Transformer-based local temporal aggregator (L-aggregator) with a window size of λ generates the local fine-grained temporal feature sequence $(l_1, ..., l_t)$. The Key-recorder is then used to record global appeared key information, where g_t represents the recorded key information from the beginning to the t-th frame. Last, a Fusion head is adopted to fuse l_t and g_t to predict the phase $\hat{p}_t$ while a linear layer uses l_t to predict the current phase transition map value $\hat{h}_t$. Note that we only need to run $\mathcal{S}^{\mathcal{R}}$ and L-aggregator once during inference. **Fusion head.** A linear layer first embeds the first branch (black) input into the same length as the second branch (grey), followed by element-wise addition, and a residual layer [18] that is tailed to output final fusion. **L-aggregator.** It contains an m-layer self-attention for encoding the first branch, and an n-layer cascaded self-attention and cross-attention decoder.

works [14, 5, 6] attempted to first train a spatial feature extractor and then freeze its weights before training the rest of the temporal modules. However, learning phases from a single image frame can be challenging since the feature extractor model may have similar image frames appearing at different phases as inputs, and phase recognition depends not only on the current frame but also on previous frames. To address these problems, we use the method proposed in [29] to train a temporally-rich spatial feature model $S^{\mathcal{R}}$ which is based on ViT [9]. We input a sequence of 30 image frames $X'_t \subseteq X_t = \{x_i\}_{i=1}^{t}$ where x_i is the i-th frame of online video stream X_t, and t is the current frame index. X'_t consists of image frames sampled at equal intervals from the beginning of the current phase up to the current frame, and is used to build a more robust phase recognition compared to the one that uses only one single image frame. We first use $\mathcal{S}^{\mathcal{R}}$ to embed all frames of X'_t into spatial features independently, followed up by a Transformer that aggregates them to predict the phase $\hat{p}_t$ of the current image frame x_t. The temporally-rich supervision method $\mathcal{S}^{\mathcal{R}}$ can be more well-trained and could extract spatial features more accurately. Note that, for each x_i we define the extracted spatial feature by $\mathcal{S}^{\mathcal{R}}$ as e_i, and we freeze the weight of $\mathcal{S}^{\mathcal{R}}$ while training the temporal module of SKiT during the second stage.

3.2. Local temporal feature aggregator

The spatial features extracted by $\mathcal{S}^{\mathcal{R}}$ are used by a Transformer-based local temporal feature aggregator *L-aggregator* that captures the current fine-grained temporal information, such as actions, moving tools, and operated organs. L-aggregator is defined as follows:

$$(l_{t-\lambda+1}, ..., l_t) = \textit{L-aggregator}\,(e_{t-\lambda+1}, ..., e_t). \quad (1)$$

Here, e_t is the current spatial feature extracted by $\mathcal{S}^{\mathcal{R}}$, l_t represents the current local temporal feature of the t-th frame, and λ is the input window size of the L-aggregator.

We implement L-aggregator using a Transformer-based module inspired by [44, 29], as shown in Figure 2. It consists of an m-layer self-attention encoder that receives one branch input and an n-layer cascaded self-attention and cross-attention decoder that receives the second branch

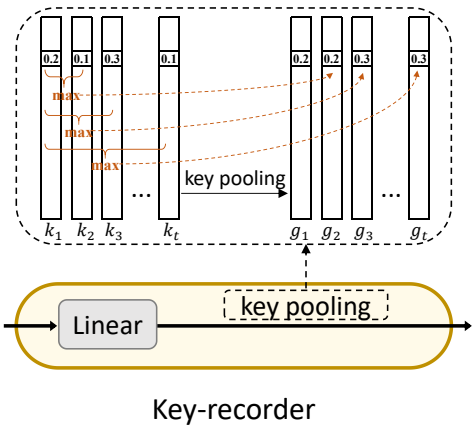


Figure 3: Key-recorder module. A linear layer embeds inputs into low-dimensional key features, followed by our key pooling operation to record the appeared key information for every current frame. We illustrate the working principle of the key pooling with a t-length key feature sequence $(k_1, ..., k_t)$, whereby for each element position, we take the maximum value from the beginning of the sequence up to the current element along the sequence direction (see arrows in brown) to get a global feature sequence $(g_1, ..., g_t)$.

input and fuses it with the encoder's output. We then duplicate the current λ-length spatial feature sequence $(e_{t-\lambda+1}, ..., e_t)$ into two branches and input them into the decoder and encoder, respectively. Note that, we use sliding windows to promise an online setting during the training stage.

3.3. Global Key-recorder

Global key information and its long-term dependencies across surgical phases are crucial for surgical phase recognition. While previous methods have proposed the integration of global features using complex temporal models, inference time increases with input sequence length, even when using efficient self-attention mechanisms such as *ProbSparse* [44], which is used by LoViT [29]. Moreover, these complex temporal models are easy to over-fit and result in remembering confusing information. To address these issues, we propose *Key-recorder*, which uses **key pooling** to only record the limited appeared key information, and discard confusing information.

As shown in Figure 3, given a local temporal feature sequence $L_t = (l_1, l_2, ..., l_t)$, we first use a linear layer to embed them into a low-dimensional key feature sequence $K_t = \{k_i\}_{i=1}^{t}$, where $k_i \in \mathbb{R}^{d_k}$ is the i-th feature vector of K_t containing d_k key information. The j-th value of k_i represents the response level of the i-th frame local feature to the j-th key information. In other words, it represents the possibility that the i-th frame contains the j-th key information, where the key information could be the result of a feature with actual physical meaning, such as tool, organ, and action, or other abstract meaning, which is learned by the model. Key pooling then aims to determine if the j-th key information appears in the global sequence up to the i-th frame. Specifically, Key pooling uses an element-wise max operation to record the maximum value of the i-th position in feature vectors from the beginning of the feature sequence up to the j-th feature vector as the value g_i^j:

$$g_i^j = \begin{cases} max(g_{i-1}^j, k_i^j), & \text{if } i > 1 \\ k_i^j, & \text{otherwise} \end{cases}. \tag{2}$$

During test inference, Key-recorder only needs to compare two feature vectors g_{t-1} and k_t to recall the appeared global key information g_t needed for recognizing the current t-th frame, with time complexity of $\mathcal{O}(1)$. As a result, the time consumption of the Key-recorder is negligible and independent of the length of the surgical video.

3.4. Fusion Head

The phase of the current frame is then determined by the current local information and the previous global information. Specifically, we adopt a small fusion head that combines the global appeared key feature g_t with the current local feature l_t. As shown in Figure 2, we first adopt a linear layer to encode the key feature g_t to the same dimension as l_t before adding them together. Following that, a residual layer [18] is adopted to optimize the output.

3.5. Loss Function

In order to give importance to the key information existing during the phase transitions, we employ the phase transition-aware supervision mechanism proposed in LoViT [29]. It consists of phase transition points projected onto a one-dimensional left-right asymmetric Gaussian kernel heatmap, called phase transition map $H = \{h_i\}_{i=1}^{T}$, where h_i is the i-th phase transition map value of T-frame video, expressed as:

$$h_i = \begin{cases} \exp(-\frac{(i-b_{p_i})^2}{2\sigma_l^2}), & b_{p_i} - 3\sigma_l < i < b_{p_i} \\ \exp(-\frac{(i-b_{p_i})^2}{2\sigma_r^2}), & b_{p_i} < i < b_{p_i} + 3\sigma_r \\ 0, & otherwise \end{cases} \tag{3}$$

where b_{p_i} is the index of the frame where the current phase p_i starts, and $3\sigma_l$ and $3\sigma_r$ are the left- and right-side kernel length departing from b_{p_i} respectively. Consequently, our loss function is a weighted sum of the phase transition map loss and phase classification loss:

$$\mathcal{L} = \mathcal{L}_1(\hat{h}, h) + \mathcal{L}_{CE}(\hat{p}, p). \tag{4}$$

Here, $\mathcal{L}_1(\hat{h}, h)$ denotes the L_1 loss between the predicted heatmap $\hat{h}$ and its ground truth h, whereas $\mathcal{L}_{CE}(\hat{p}, p)$ represents the cross-entropy loss between the predicted phase $\hat{p}$ and its ground truth p.

4. Experiments

4.1. Experimental Design

Datesets. We conducted experiments on two public surgical video datasets: Cholec80 [37] and AutoLaparo [40]. *Cholec80* comprises 80 laparoscopic surgical videos, with an average duration of 39 minutes at 25 frames per second (fps). The dataset includes manual annotations of seven surgical phases that we used for our study. We kept the dataset split into 40 videos for training and 40 videos for testing, following previous works [5, 14, 29]. *AutoLaparo* consists of 21 videos with seven phases, with an average video duration of 66 minutes recorded at 25 fps. We split the dataset into 10 videos for training, 4 videos for validation, and 7 videos for testing following [40, 29]. Both datasets were sampled into 1 fps following previous works [14, 29].

Training details. Our experiments were conducted on a single NVIDIA Tesla V100 GPU. We utilized a 12-head, 12-layer Transformer encoder as our spatial feature extractor $\mathcal{S}^{\mathcal{R}}$ based on the ViT-B/16 architecture following LoViT [29]. This model was pretrained on ImageNet-1K (IN1k) [7] and produced 768D representations, with an input image size of 248×248 pixels. We trained the feature extractor using SGD+momentum for 35 epochs, with a 5-epoch warm-up period [17] and a 30-epoch cosine annealed decay. For the local temporal feature extractor, we used a window size of $\lambda = 100$, producing 512D feature vectors with $m = 2$, $n = 2$. The Key-recorder generated 64D and 32D key information representations on Cholec80 and AutoLaparo respectively. The temporal modules were trained for 40 epochs using SGD+momentum with a learning rate of $3e{-}4$, weight decay of $1e{-}5$, a 5-epoch warm-up period [17], and a 35-epoch cosine annealed decay, with a batch size of 8.

Metrics. To assess our model's effectiveness, we use four widely-used benchmark metrics for surgical phase recognition: accuracy, precision, recall, and Jaccard. Accuracy is a video-based measure, indicating the percentage of correctly recognized phases while minimizing the effect of video length. However, the class (phase) distribution in the dataset is imbalanced, and short phases have little impact on overall video accuracy. Therefore, we additionally employ class-level precision, recall, and Jaccard to evaluate the performance of our model across different dimensions. These metrics represent the positive predictive value, positive rate, and intersection rate of recognition versus ground truth at the phase level, respectively.

Prior works [14, 22, 42] employed a relaxed metric, which lacks a clear explanation in their manuscripts. This metric considers predictions falling into neighbouring phases within a 10-second window around the phase transition as correct, even if they do not match the ground truth. Nonetheless, phase transition prediction is a vital model indicator [29].

Dataset	Method	Relaxed metric	Video-level Metric	Phase-level Metric		
			Accuracy ↑	Precision ↑	Recall ↑	Jaccard ↑
Cholec80	PhaseNet [36]	✓	78.8 ± 4.7	71.3	76.6	-
	SV-RCNet [20]	✓	85.3 ± 7.3	80.7	83.5	-
	OHFM [41]	✓	87.3 ± 5.7	-	-	67.0
	TeCNO [5]	✓	88.6 ± 7.8	86.5	87.6	75.1
	TMRNet [22]	✓	90.1 ± 7.6	90.3	89.5	79.1
	Trans-SVNet [14]	✓	90.3 ± 7.1	90.7	88.8	79.3
	Not End-to-End [42]	✓	91.5 ± 7.1	-	86.8	77.2
	LoViT [29]	✓	92.4 ± 6.3	89.9	90.6	81.2
	SKiT (ours)	✓	**93.4 ± 5.2**	**90.9**	**91.8**	**82.6**
	Trans-SVNet		89.1 ± 7.0	**84.7**	83.6	72.5
	AVT [16]		86.7 ± 7.6	77.3	82.1	66.4
	TeSTra [43]		90.1 ± 6.6	82.8	83.8	71.6
	LoViT [29]		91.5 ± 6.1	83.1	86.5	74.2
	SKiT (ours)		**92.5 ± 5.1**	84.6	**88.5**	**76.7**
AutoLaparo	SV-RCNet		75.6	64.0	59.7	47.2
	TMRNet		78.2	66.0	61.5	49.6
	TeCNO		77.3	66.9	64.6	50.7
	Trans-SVNet		78.3	64.2	62.1	50.7
	AVT		77.8	68.0	62.2	50.7
	LoViT [29]		81.4 ± 7.6	**85.1**	65.9	56.0
	SKiT (ours)		**82.9 ± 6.8**	81.8	**70.1**	**59.9**

Table 1: The results (%) of different state-of-the-art methods on both the Cholec80 and AutoLaparo datasets. The best results are marked in bold.

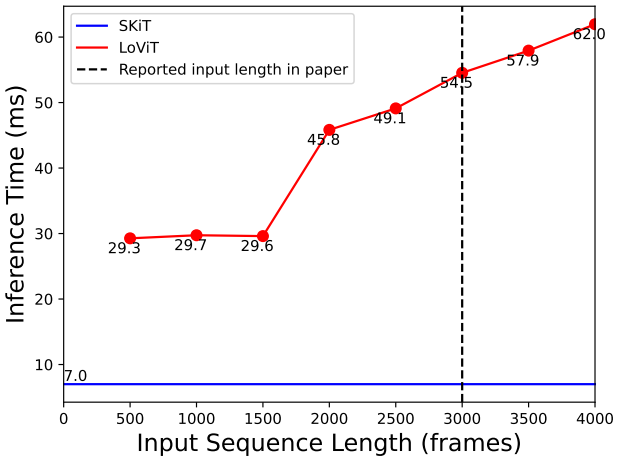
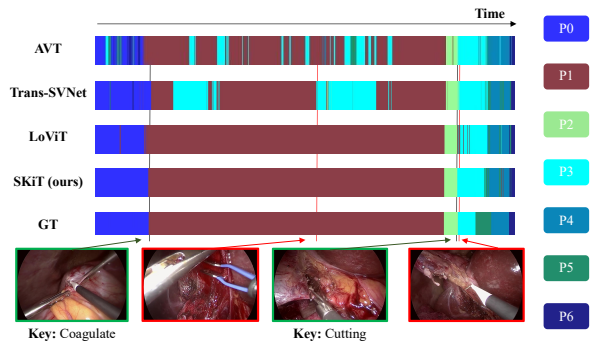


Figure 4: Inference time comparison with LoViT for different input video lengths.

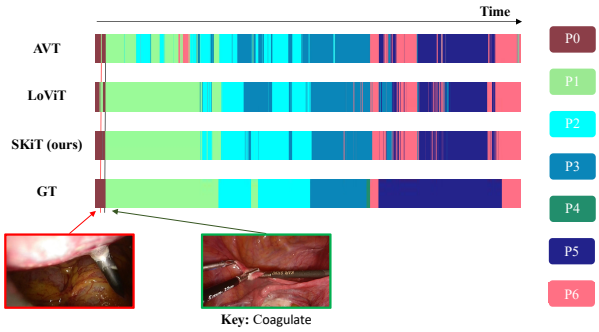
Additionally, to handle missing phases in some videos while computing the standard metric, we combined the predictions and ground truth of all videos into one sequence and computed the average per phase. However, we discovered that the previous code using the relaxed boundary-based metric calculated each phase in each video independently and then obtained the average score but overlooked 'NaN', which was not introduced before, and We kept the same approach while using the relaxed boundary metric.

4.2. Comparison With State-of-the-art Methods

We conducted a comparative study of SKiT with other state-of-the-art methods for surgical phase recognition and action anticipation on Cholec80 and AutoLaparo datasets.



(a) Phase Recognition predictions on Cholec80.



(b) Phase Recognition predictions on AutoLaparo.

Figure 5: Qualitative comparisons with other methods on selected frames from Cholec80 and AutoLaparo datasets. We show multiple methods' results on a given video with its ground truth (GT). Our method is able to correctly recognise key transition frames (green) and ignore ambiguous frames (red), whereas previous methods fail in those specific misleading cases.

Table 1 shows the quantitative results. SKiT outperformed the previous state-of-the-art method, LoViT, with an improvement in accuracy of 1 pp (percentage points) on Cholec80 (from 91.5% to 92.5%) and 1.5 pp on AutoLaparo (from 81.4% to 82.9%), while maintaining a simpler and more efficient framework. Moreover, SKiT exhibited superior phase-level metrics such as precision, recall, and Jaccard, which address the issue of phase imbalance effects. SKiT also achieved lower standard deviations of accuracy across different videos, with a reduction of 1 pp and 0.8 pp on Cholec80 and AutoLaparo datasets, respectively, compared to LoViT. This indicates that SKiT achieved more stable performance across various surgical videos. Following the same setup as Trans-SVNet on the Cholec80 dataset, we do not have a separate validation set. However, the trained model of the final epoch shows only a slight 0.3 pp decrease in accuracy compared to the reported best model's performance, which proves that our model is not overfitting. It is important to note that some methods, such as Opera [6], did not make their code publicly available and used different approaches to split the training and testing datasets. Therefore, we were unable to compare our results with these methods. For the Anticipative Video Transformer (AVT) [16] method, which was proposed for action anticipation, we implemented it using the released code with 32-frame input. Moreover, we also implemented TeSTra [43] using their official code. As for the results on the AutoLaparo dataset, we used the reported results on [40]. It is worth mentioning that most of the reported methods were from the same team that worked on this dataset, which adds to the credibility of the reported results. We present representative visual results for phase recognition on Cholec80 and AutoLaparo datasets in Figure 5. Overall, SKiT outperformed other methods in most places where "key feature" (as shown in frames highlighted in green) was recognized by our model as key information to correctly classify the phase, while being robust to confusing image frames.

4.3. Runtime Analysis

We study the inference time of SKiT in contrast with the previous version of LoViT using different numbers of input frames, as illustrated in Figure 4. It should be noted that the measured values may be slower than the actual values due to testing on a GPU cluster and sharing the CPU with other tasks, which leads to unstable inference times. To compare the temporal module inference time, we excluded the spatial feature extractor inference time, which takes about 9 ms per frame of inference time. Our results show that SKiT is *eight times faster* than LoViT, with an online inference time of approximately 7 ms per frame (142 fps) *versus* 55 ms per frame (18 fps) for the reported 3000 frames input in [29]. Notably, the speed advantage of SKiT increases as the input length becomes longer. Additionally, the inference time of SKiT is not dependent on the input length required to capture the entire video-length receptive field, making it highly suitable for constructing a stable and efficient surgical recognition system. Although Trans-SVNet achieved a high-speed inference time of 10 ms per frame (91 fps) reported in [14], its perceptive field is fixed after design due to the TCN structure, which means it cannot maintain the whole-video receptive field when processing videos of variable lengths. TMRNet [22] has the ability to process variable-length videos while maintaining a global receptive field with the non-local structure, but its inference time increases with the length of the video, and its average inference speed is slow, at 80 ms per frame (11 fps) as reported in the original paper.

4.4. Ablation Study

Key-recorder module. We investigated the impact of our proposed Key-recorder for recording globally appeared

Feature	Video-level Metric	Phase-level Metric		
	Accuracy	Precision	Recall	Jaccard
l	89.3 ± 7.3	81.1	85.1	71.1
$g^{\mathcal{T}}$	90.9 ± 6.9	82.1	88.1	74.2
$l + g^{\mathcal{S}}$	88.1 ± 7.1	77.8	80.8	65.9
$l + g^{\mathcal{T}}$	$\mathbf{92.5 \pm 5.1}$	**84.6**	**88.5**	**76.7**

Table 2: The results (%) of different combinations of features on both the Cholec80 dataset. Combination of features include: (l) - using only local temporal feature sequence extracted from *L-aggregator*, ($g^{\mathcal{T}}$) - using only the global key feature recorded from local features l, ($l + g^{\mathcal{S}}$) - l in combination with the global key feature recorded from spatial features e, and ($l + g^{\mathcal{T}}$) - l in combination with the global key feature recorded from local features l. The best results are marked in bold.

key information. Table 2 presents the results, where 'l' represents the local temporal feature extracted by the L-aggregator, '$g^{\mathcal{S}}$' denotes the global key feature recorded from spatial feature e, and '$g^{\mathcal{T}}$' stands for the key feature recorded from local temporal feature l. By adding the global key feature recorded from the local temporal feature, we achieved improvements in recognition performance on both datasets across all metrics. For example, on the Cholec80 dataset, we observed an accuracy improvement of 3.2 pp (from 89.3% to 92.5%). This highlights the significance of capturing global appeared key information for recognizing the current phase. We observed that if the global feature is solely recorded from the spatial feature e, the resulting recorded key feature $g^{\mathcal{S}}$ may have a negative impact on phase recognition. This is because the spatial feature alone may not accurately capture the status of every frame, including some ambiguous frames, leading to noise in the recorded key feature. This noise negatively affects the overall quality of the recorded key information since it is considered in the maximum operation of our key pooling mechanism. To address this issue, we incorporated local window size frames to more accurately describe the features of each frame and reduce the negative impact of noise.

We also replace our Key-recorder with some other temporal modules and do some experiments on Cholec80 as shown in Table 3, it shows that the Key-recorder outperforms the rest three classic backbones, LSTM [19], GRU [4], and TCN [26], which further proves the effeteness of the proposed Key-recorder.

Key feature length. We also examined the impact of the length parameter d_g on the recognition performance of Key-recorder. The length of the key feature d_g determines the amount of key information that Key-recorder can store. As shown in Table 4, the model's performance is affected by this length parameter, highlighting the importance of storing an appropriate amount of key information to improve

Method	Video-level Metric	Phase-level Metric		
	Accuracy	Precision	Recall	Jaccard
L-aggregator + LSTM	89.5 ± 7.5	80.8	86.5	71.8
L-aggregator + GRU	89.6 ± 6.8	81.7	84.1	70.8
L-aggregator + TCN	90.1 ± 7.2	82.0	85.8	72.5
L-aggregator + Key-recorder (SKiT)	$\mathbf{92.5 \pm 5.1}$	**84.6**	**88.5**	**76.7**

Table 3: The results (%) of different temporal modules replacing with our Key-recorder of SKiT on Cholec80 dataset. The best results are marked in bold.

Dataset	d_k	Video-level Metric	Phase-level Metric		
		Accuracy	Precision	Recall	Jaccard
Cholec80	8	90.6 ± 6.4	82.9	86.2	73.6
	16	90.6 ± 6.8	83.3	86.3	74.0
	32	92.1 ± 5.4	**85.1**	86.6	75.7
	64	$\mathbf{92.5 \pm 5.1}$	84.6	**88.5**	**76.7**
	128	92.1 ± 5.6	83.8	88.1	75.7
AutoLaparo	8	82.6 ± 6.9	79.6	**70.1**	**60.5**
	16	80.9 ± 8.4	73.0	64.7	55.0
	32	$\mathbf{82.9 \pm 6.8}$	81.8	**70.1**	59.9
	64	81.9 ± 6.9	**82.7**	68.9	58.4
	128	81.5 ± 8.2	70.3	66.1	55.5

Table 4: The results (%) of different key feature lengths, d_k, with SKiT on both the Cholec80 and the AutoLaparo datasets. The best results are marked in bold.

recognition performance. Insufficient key information can lead to inadequate learning, while excessive information can result in redundancy and overfitting. Therefore, it is crucial to strike a balance between the amount of key information and the risk of overfitting. In other words, Key-recorder should store a limited number of key events that are likely to occur in the video to efficiently capture the necessary information.

λ	Video-level Metric	Phase-level Metric		
	Accuracy	Precision	Recall	Jaccard
20	$\mathbf{92.5 \pm 5.3}$	84.3	88.4	76.4
50	92.4 ± 5.1	**84.9**	87.3	76.1
100	$\mathbf{92.5 \pm 5.1}$	84.6	**88.5**	**76.7**
200	92.3 ± 5.0	84.5	87.8	76.2

Table 5: The results (%) of different local temporal window sizes, λ, with SKiT on Cholec80 dataset. The best results are marked in bold.

Local size. We conducted a comparison of different local sizes λ for the L-aggregator on the Cholec80 dataset, as presented in Table 5. Our results indicate that the choice of λ does not significantly impact the proposed SKiT.

Different Input Sizes. We experimented with different input sizes T, as detailed in Table 6, and found that longer inputs generally led to better performance. In order to train SKiT using multi-batch sizes, we did not use the entire video as input.

T	Video-level Metric	Phase-level Metric		
	Accuracy	Precision	Recall	Jaccard
100	90.2 ± 7.0	82.3	86.2	73.1
500	91.9 ± 5.4	84.5	87.5	76.0
1000	92.0 ± 5.0	**85.0**	86.6	75.6
3000	$\mathbf{92.5 \pm 5.1}$	84.6	**88.5**	**76.7**

Table 6: The results (%) of different input sizes T of SKiT on the Cholec80 dataset. The best results are marked in bold.

5. Visualization of Key Information

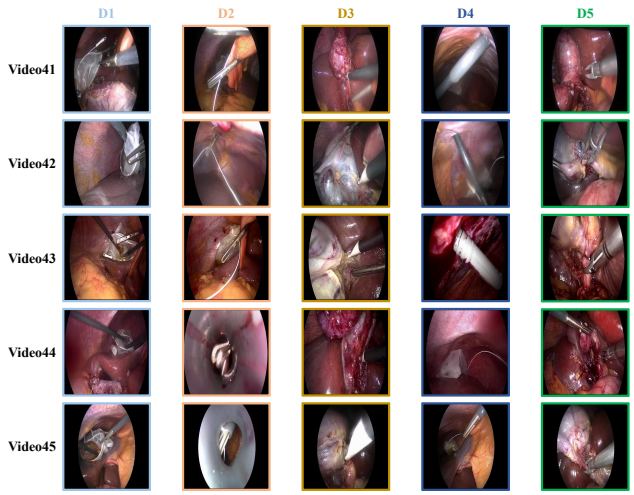


Figure 6: **Illustrative examples of key information.** Five videos from the Cholec80 test dataset (rows) are used to visualise the key feature $k \in \mathbb{R}^{d_k}$ resulting from our local temporal feature l. We randomly choose five dimensions D1, D2, D3, D4, D5 (columns) out of the total of 32 dimensions ($d_k = 32$) and plot the image frames corresponding to the maximum value of the key feature along the entire video. We highlight that the key information, represented as each dimension, has similarities across videos demonstrating the interpretability of key events occurring in videos.

In order to gain insight into the key information presented in each dimension of the key feature, we plotted the image frames that corresponded to the maximum value of each dimension in key feature k throughout the entire video. Some examples of these plots are shown in Figure 6. It is important to note that each dimension of the key feature carries distinct physical implications, which we refer to as key information, supporting our initial assumption. For example, we observed that all instances in the fifth dimension (D5) include the tool 'Clipper'. This finding confirms the relevance of the key feature and highlights its potential to capture important features in surgical videos.

6. Conclusion

We propose a fast and effective method for surgical phase recognition called SKiT, which utilizes a novel Key-recorder to record limited key information that appears in surgical videos. Our proposed method achieves improved performance compared to previous state-of-the-art work, while having inference times of temporal models approximately *eight* times faster than LoViT. The time complexity of the key pooling operation as part of the Key-recorder is only $\mathcal{O}(1)$, making it fast and enabling the inference time of SKiT to remain unaffected by the number of input video frames. However, despite achieving state-of-the-art performance with efficient inference time, SKiT still requires a two-step training process due to the lengthy nature of surgical videos, which is a limitation existing in other surgical phase recognition methods. While we evaluated the performance of Key-recorder on different datasets and through ablation studies in addressing some of the limitations of the previous state-of-the-art work [29], further investigation is necessary to ascertain the effect Key-recorder has on other proposed models. Moreover, although the proposed model can reuse previous global recorded key feature while recognising the current frame to greatly improve inference time, it remains difficult to do so during training resulting in a process that is still time-consuming. If some information can be reused in training and the training cost can be reduced, end-to-end training of the model with long sequence inputs could be achieved, a topic that is in our future research plans.

Acknowledgement

This work was supported by the Engineering & Physical Sciences Research Council Doctoral Training Partnership (EPSRC DTP) grant EP/T517963/1; the Academy of Medical Sciences Springboard Award [SBF005\1131]; King's funded CDT in Surgical & Interventional Engineering and King's-China Scholarship Council PhD Scholarship programme (K-CSC).

References

[1] Seyed-Ahmad Ahmadi, Tobias Sielhorst, and Nassir Navab. Recovery of surgical workflow without explicit models. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2006, 9th International Conference, Copenhagen, Denmark, October 1-6, 2006, Proceedings, Part I*, pages 420–428, 2006. 1

[2] Tobias Blum, Nicolas Padoy, and Nassir Navab. Workflow mining for visualization and analysis of surgeries. *Int. J. Comput. Assist. Radiol. Surg.*, 3(5):379–386, 2008. 1

[3] Loubna Bouarfa, Pieter P. Jonker, and Jenny Dankelman. Discovery of high-level tasks in the operating room. *J. Biomed. Informatics*, 44(3):455–462, 2011. 1

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 8

[5] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part III*, pages 343–352, 2020. 3, 4, 6

[6] Tobias Czempiel, Magdalini Paschali, Daniel Ostler, Seong Tae Kim, Benjamin Busam, and Nassir Navab. Opera: Attention-regularized transformers for surgical phase recognition. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part IV*, pages 604–614, 2021. 3, 4, 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009. 6

[8] Olga Dergachyova, David Bouget, Arnaud Huaulmé, Xavier Morandi, and Pierre Jannin. Automatic data-driven real-time segmentation and recognition of surgical workflow. *Int. J. Comput. Assist. Radiol. Surg.*, 11(6):1081–1089, 2016. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 4

[10] Haoqi Fan, Bo Xiong, and Christoph Feichtenhofer. Multi-scale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 3

[11] Yazan Abu Farha and Jürgen Gall. MS-TCN: multi-stage temporal convolutional network for action segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3575–3584, 2019. 3

[12] Stefan Franke, Max Rockstroh, and Thomas Neumuth. The intelligent OR: design and validation of a context-aware surgical working environment. *Int. J. Comput. Assist. Radiol. Surg.*, 13(8):1301–1308, 2018. 1

[13] Xiaojie Gao, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 8440–8446, 2020. 3

[14] Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part IV*, pages 593–603, 2021. 3, 4, 6, 7

[15] Carly R Garrow, Karl-Friedrich Kowalewski, Linhong Li, Martin Wagner, Mona W Schmidt, Sandy Engelhardt, Daniel A Hashimoto, Hannes G Kenngott, Sebastian Bodenstedt, Stefanie Speidel, et al. Machine learning for surgical phase recognition: a systematic review. *Annals of surgery*, 273(4):684–693, 2021. 1

[16] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 13485–13495, 2021. 3, 6, 7

[17] Priya Goyal, Piotr Dollár, Ross B. Girshick, and Kaiming He. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017. 6

[18] Kaiming He, Xiangyu Zhang, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3, 4, 5

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 3, 8

[20] Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network. *IEEE Trans. Medical Imaging*, 37(5):1114–1126, 2018. 3, 6

[21] Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical Image Anal.*, 59, 2020. 1, 2, 3

[22] Yueming Jin, Yonghao Long, and Pheng-Ann Heng. Temporal memory relation network for workflow recognition from surgical video. *IEEE Trans. Medical Imaging*, 40(7):1911–1923, 2021. 3, 6, 7

[23] Darko Katic, Anna-Laura Wekerle, and Stefanie Speidel. Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance. In *Information Processing in Computer-Assisted Interventions - 5th International Conference, IPCAI 2014, Fukuoka, Japan, June 28, 2014. Proceedings*, pages 158–167, 2014. 1

[24] Karl-Friedrich Kowalewski and Felix Nickel. Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. *Surgical endoscopy*, 33:3732–3740, 2019. 1

[25] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. *Int. J. Comput. Vis.*, 128(3):642–656, 2020. 3

[26] Colin Lea, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, pages 47–54, 2016. 3, 8

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[28] Yanghao Li, Chao-Yuan Wu, , and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 3

[29] Yang Liu, Maxence Boels, Luis C Garcia-Peraza-Herrera, Tom Vercauteren, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. Lovit: Long video transformer for surgical phase recognition. *arXiv preprint arXiv:2305.08989*, 2023. 3, 4, 5, 6, 7, 9

[30] Lena Maier-Hein, Matthias Eisenmann, and Stefanie Speidel. Surgical data science - from concepts to clinical translation. *CoRR*, abs/2011.02284, 2020. 1

[31] Hirenkumar Nakawala, Roberto Bianchi, and Elena De Momi. "deep-onto" network for surgical workflow and context recognition. *Int. J. Comput. Assist. Radiol. Surg.*, 14(4):685–696, 2019. 3

[32] Thomas Neumuth. Surgical process modeling. *Innovative surgical sciences*, 2(3):123–137, 2017. 1

[33] Gwenolé Quellec, Mathieu Lamard, and Guy Cazuguel. Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials. *IEEE Trans. Medical Imaging*, 34(4):877–887, 2015. 1

[34] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. 1

[35] Andru Putra Twinanda. *Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos. (Approches basées vision pour la reconnaissance d'activités chirurgicales à partir de vidéos laparoscopiques et multi-vues RGBD)*. PhD thesis, University of Strasbourg, France, 2017. 3

[36] Andru Putra Twinanda, Didier Mutter, and Nicolas Padoy. Single- and multi-task architectures for surgical workflow challenge at M2CAI 2016. *CoRR*, abs/1610.08844, 2016. 6

[37] Andru Putra Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Medical Imaging*, 36(1):86–97, 2017. 1, 2, 6

[38] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125, 2016. 3

[39] Ashish Vaswani, Noam Shazeer, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 1, 3

[40] Ziyi Wang, Bo Lu, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*, 2022. 6, 7

[41] Fangqiu Yi and Tingting Jiang. Hard frame detection and online mapping for surgical phase recognition. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part V*, pages 449–457, 2019. 3, 6

[42] Fangqiu Yi, Yanfeng Yang, and Tingting Jiang. Not end-to-end: Explore multi-stage architecture for online surgical phase recognition. In *Computer Vision - ACCV 2022 - 16th Asian Conference on Computer Vision, Macao, China, December 4-8, 2022, Proceedings, Part IV*, volume 13844, pages 417–432. Springer, 2022. 6

[43] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision*, pages 485–502. Springer, 2022. 3, 6, 7

[44] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 11106–11115, 2021. 2, 3, 4, 5

[45] Odysseas Zisimopoulos, Evangello Flouty, and Danail Stoyanov. Deepphase: Surgical phase recognition in CATARACTS videos. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV*, volume 11073 of *Lecture Notes in Computer Science*, pages 265–272. Springer, 2018. 3