Video Summarization With Spatiotemporal Vision Transformer

Tzu-Chun Hsu, Yi-Sheng Liao, and Chun-Rong Huang, *Senior Member, IEEE*

***Abstract*— Video summarization aims to generate a compact summary of the original video for efficient video browsing. To provide video summaries which are consistent with the human perception and contain important content, supervised learning-based video summarization methods are proposed. These methods aim to learn important content based on continuous frame information of human-created summaries. However, simultaneously considering both of inter-frame correlations among non-adjacent frames and intra-frame attention which attracts the humans for frame importance representations are rarely discussed in recent methods. To address these issues, we propose a novel transformer-based method named spatiotemporal vision transformer (STVT) for video summarization. The STVT is composed of three dominant components including the embedded sequence module, temporal inter-frame attention (TIA) encoder, and spatial intra-frame attention (SIA) encoder. The embedded sequence module generates the embedded sequence by fusing the frame embedding, index embedding and segment class embedding to represent the frames. The temporal inter-frame correlations among non-adjacent frames are learned by the TIA encoder with the multi-head self-attention scheme. Then, the spatial intra-frame attention of each frame is learned by the SIA encoder. Finally, a multi-frame loss is computed to drive the learning of the network in an end-to-end trainable manner. By simultaneously using both inter-frame and intra-frame information, our method outperforms state-of-the-art methods in both of the SumMe and TVSum datasets. The source code of the spatiotemporal vision transformer will be available at https://github.com/nchucvml/STVT.**

***Index Terms*— Video summarization, transformer, vision transformer, multi-head self-attention, temporal inter-frame correlation, spatial intra-frame attention, multi-frame loss.**

I. Introduction

With the increasing number of videos in the internet, efficiently browsing videos becomes one of the most important issues in the video processing domain. To address this issue, video summarization methods [1], [2], [3] are proposed and have been shown their effectiveness. These methods can also be applied to various applications including video saliency analysis [4], [5], video synopsis [6], [7] and video content analysis [8], [9].

Video summarization can be achieved by using unsupervised methods [10], [11], [12], weakly supervised methods [13], [14], [15], and supervised methods [16], [17], [18]. While unsupervised methods extract keyframes based on frame similarity, weakly supervised methods further exploit auxiliary information to help identify keyframes for video summarization. Nevertheless, the semantic concept of the video content which attracts the humans' attentions may not be correctly represented by using frame similarity. Thus, unsupervised and weakly supervised methods [19], [20] are hard to generate summaries which are consistent with the human perception. To solve the aforementioned problem, supervised methods [16], [17], [18], [19], [20], [21] are proposed to learn the frame importance from human-created summaries.

Recently, long short-term memory (LSTM)-based video summarization methods [17] are proposed. Based on the continuous frame correlations learned by LSTM, these methods achieve frame-level importance prediction and have been shown more effective compared with conventional methods. However, most LSTM-based methods mainly consider temporal correlations between adjacent frames, i.e. these methods only learn continuous contextual information. Inter-frame correlations among non-adjacent frames are not well learned by using LSTM. Moreover, intra-frame content which can attract the human's attention is not considered. As indicated in [20] and [22], the LSTM-based methods are hard to well model long-range dependency among video frames and usually apply down-sampling to the training data [17], [19], [20], [21]. The frames of short shots containing important contextual information may be discarded due to down-sampling. In addition, the gradient vanishing problem usually occurs for these methods as indicated in [22] and [23]. To solve these problems, attention-based video summarization methods [22], [24], [25], [26] are proposed. These methods exploit temporal frame dependency by using attention schemes. Nevertheless, these methods are hard to represent relations among frames and spatial information within frames [18]. As a result, developing a novel deep learning network which can effectively learn the frame importance based on spatial and temporal information of video summaries becomes one of the most challenging and important issues in video summarization.

To solve the aforementioned problems, we propose a novel transformer-based method named spatiotemporal vision

Manuscript received 12 December 2021; revised 18 October 2022 and 5 April 2023; accepted 27 April 2023. Date of publication 15 May 2023; date of current version 26 May 2023. This work was supported in part by the National Science and Technology Council of Taiwan under Grant NSTC 111-2634-F-006-012, Grant NSTC 111-2628-E-006-011-MY3, and Grant NSTC 112-2327-B-006-008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Francesco G. B. De Natale. *(Corresponding author: Chun-Rong Huang.)*

Tzu-Chun Hsu and Yi-Sheng Liao are with the Department of Computer Science and Engineering, National Chung Hsing University, Taichung 402, Taiwan (e-mail: g109056029@nchu.edu.tw; g107056049@nchu.edu.tw).

Chun-Rong Huang is with the Cross College Elite Program and the Academy of Innovative Semiconductor and Sustainable Manufacturing, National Cheng Kung University, Tainan 701, Taiwan, and also with the Department of Computer Science and Engineering, National Chung Hsing University, Taichung 402, Taiwan (e-mail: crhuang@gs.ncku.edu.tw).

Digital Object Identifier 10.1109/TIP.2023.3275069

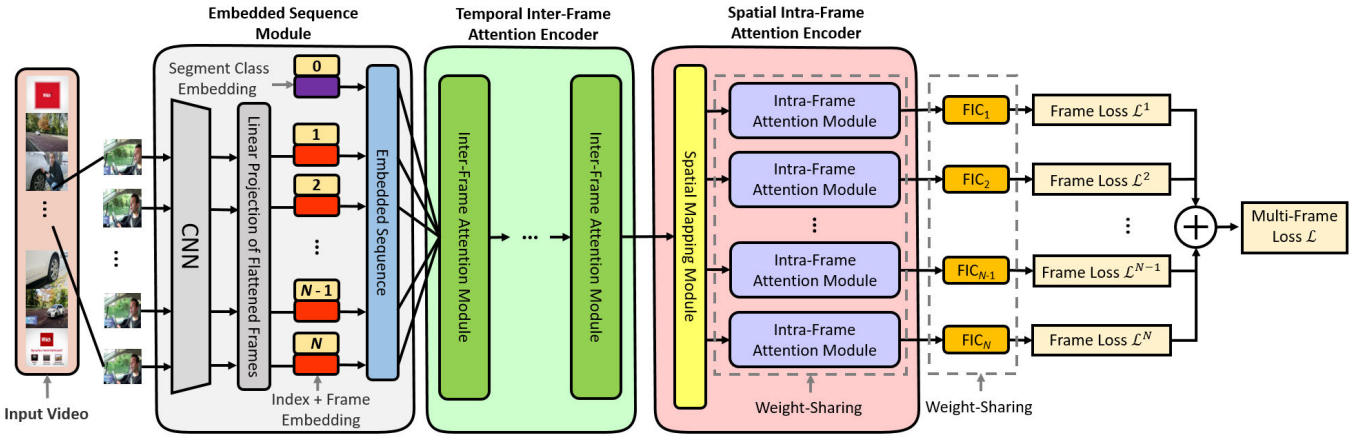


Fig. 1. The overview of the spatiotemporal vision transformer (STVT) for video summarization. In the embedded sequence module, the deep frame features of each frame are extracted by using a CNN network and are projected by a trainable frame embedding projection to the frame embedding. The frame embedding is combined with the segment class embedding and index embedding to generate the embedded sequence, which serves as the input of the TIA encoder. The TIA encoder aims to learn temporal inter-frame correlations based on the multi-head self-attention schemes of the cascaded inter-frame attention modules. To further extract spatial intra-frame attention for each frame, the SIA encoder is proposed. The features learned by the SIA encoder serve as the inputs of the frame importance classifiers (FIC) for frame importance classification and frame loss computation. To drive the learning of the network in an end-to-end trainable manner, the frame losses are combined to the multi-frame loss. The weight-sharing of the modules aims to reduce the memory usage and help avoid overfitting.

transformer (STVT) for video summarization as shown in Fig. 1. Our method contains three dominant components including the embedded sequence module, temporal inter-frame attention (TIA) encoder, and spatial intra-frame attention (SIA) encoder. The embedded sequence module aims to generate the embedded sequence with the frame embedding, index embedding and segment class embedding to represent the frames. The embedded sequence then serves as the feature representation of the input video segment. Instead of learning temporal correlations between adjacent frames as in most LSTM-based methods, the proposed TIA encoder further learns temporal inter-frame correlations from the embedded sequence based on the multi-layer inter-frame attention modules with multi-head self-attention [27], [28] among frames. The long-term contextual information from video segments can be represented by our TIA encoder during learning the frame importance from the full set of the video frames.

While the TIA encoder learns the temporal inter-frame correlations based on the inter-frame attention module, each frame may also contain salient regions which attract the human's attention. To further discover the spatial intra-frame attention from each frame, the SIA encoder is proposed. The features produced by the TIA encoder are mapped by using the spatial mapping module. The mapped features are forwarded to parallel intra-frame attention modules of the SIA encoder for multi-head self-attention computation to discover the spatial attention feature of each frame. Each spatial attention feature of the frame serves as the input of the frame importance classifier (FIC) for frame importance classification. Compared with attention-based methods, the temporal inter-frame correlations learned by the TIA encoder and spatial inter-frame attention learned by the SIA encoder provide more robust spatiotemporal feature representations for human-created summaries.

To drive the learning of the network in an end-to-end trainable manner, a novel multi-frame loss is proposed based on the frame losses computed from each frame. As a result, the proposed novel transformer-based video summarization method can successfully learn the human-created summaries compared with the state-of-the-art LSTM-based and attention-based methods in both of the SumMe [29] and TVSum [30] datasets.

In addition, there are several differences between the proposed method and the preliminary method [31]. First, the proposed method can achieve the frame-level importance learning based on the multi-frame loss, while the preliminary method can only achieve the segment-level importance learning. Second, the proposed method combines the TIA encoder with the SIA encoder to learn both of the temporal inter-frame correlations among frames and spatial intra-frame attention within each frame, while the preliminary method only learns the temporal inter-frame correlations among frames. Third, an additional deep CNN is applied to generate deep frame features to boost the feature representation ability of frames in the proposed method. These differences make the proposed method achieve significantly better performance compared with the preliminary method.

The contribution of this paper is three-fold.

- To the best of our knowledge, the proposed method including the preliminary version is the first transformer-based video summarization method which introduces the concept of applying continuous temporal frames to transformer structures to describe spatiotemporal correlations between and within frames for video summarization.
- By combining the temporal inter-frame correlations learned by the TIA encoder and the spatial intra-frame attention within frames learned by the SIA encoder, our spatiotemporal vision transformer successfully represents

important temporal and spatial content compared with attention-based methods.
- The proposed transformer-based structure can effectively learn the spatiotemporal correlations from the full set of video summaries to achieve state-of-the-art performance.

The paper is organized as follows. Sec. II gives the related work of video summarization methods. The proposed method is presented in Sec. III. Sec. IV shows the experimental results and comparisons. Finally, the conclusions are given in Sec. V.

II. Related Work

Video summarization methods aim to provide informative subsets of videos by using keyframes [1], [32], [33] or keyshots [34], [35] for effective and efficient video browsing. These methods can be mainly divided into three categories, unsupervised methods, weakly-supervised methods and supervised methods, and are reviewed in the following.

A. Unsupervised Methods

To avoid the human labeling efforts, unsupervised video summarization methods are proposed. Hand-crafted features are usually used to represent video frames for frame similarity assessment. Then, video summarization is achieved by using clustering methods including [1], [33], [36], [37]. Besides clustering methods, sparse coding-based methods [2], [11], [38], [39], [40], [41] are proposed to extract compact representations based on hand-crafted features. User attention-based methods [42], [43] are also proposed to obtain attractive frames for video summarization.

With the development of the deep learning methods, generative adversarial networks (GAN) [44] are applied to learn a video summary by considering the reconstruction of the original video. For example, Mahasseni et al. [45] propose an adversarial LSTM network consisting of the summarizer and discriminator for video summarization. While the summarizer obtains the summarization results, the discriminator, which is a binary sequence classifier LSTM, is used to classify the generated video as the original video or the summary video. Yuan et al. [46] consider to maximize the mutual information between the original video and the summary video by using a cycle-consistent adversarial network. Jung et al. [12] propose using a variational autoencoder (VAE) with the GAN structure to learn frame features and assess frame dependence for keyframe selection by using attention. The further improvement of assessing frame dependence by using self-attention is proposed in [47]. Apostolidis et al. [48] combine GAN with the actor-critic model and formulate a sequence generation task to select important video fragments. Instead of applying GAN, Zhang et al. [49] propose an online motion LSTM auto-encoder with the online dictionary learning to memorize and track key object motions for unsupervised object-level video summarization.

B. Weakly-Supervised Methods

While unsupervised summarization methods consider frame feature dependence to obtain video summaries, weakly supervised video summarization methods learn to obtain video summaries from weak labels. For example, web priors [13], [50], video titles [30] and video categories [3] are served as the weak labels. More recently, deep learning-based weakly-supervised methods are proposed. Panda et al. [14] propose using a 3D convolutional neural network architecture to learn importance from the video-level annotations of different categories of web videos. Cai et al. [15] propose a LSTM-based variational encoder-summarizer-decoder (VESD) to identify important frames from web videos. Chen et al. [51] propose a weakly supervised reinforcement learning framework which is trained by subtasks with task-level binary labels and selects key shots based on rewards. Although the auxiliary information helps the learning of video importance for video summarization, the generated video summarization results are still hard to achieve comparable performance compared with supervised methods [52].

C. Supervised Methods

Supervised methods aim to discover underlying criteria for video summarization by learning from datasets with human-created summaries [52]. For example, Gygli et al. [53] propose using jointly optimizing multiple objective functions to learn the importance of a summary. Gong et al. [54] propose the sequential determinantal point process (SeqDPP) which models the process to sequentially select diverse subsets for video summarization. To further learn the local diversity of videos, Li et al. [55] propose the dynamic SeqDpp (DySeqDpp) to improve [54] by dynamically controlling the time span of a segment and learn the summaries by using a reinforcement learning algorithm. Sharghi et al. [56] apply the large-margin algorithm to the SeqDpp [54] and propose a new probabilistic block for SeqDpp to control the length of the summary. Zhang et al. [57] propose a non-parametric learning method to transfer summary structures implied by the subset selection based on training videos. Li et al. [58] design four models based on importance, representativeness, diversity and storyness to obtain the characteristics of video summaries. Instead of only considering the video content, Sharghi et al. [59] consider the user query with video content for video summarization. They propose the sequential and hierarchical determinantal point process (SH-DPP) to select keyshots based on the relevance to the user query and frame importance.

These conventional supervised methods usually apply hand-crafted features to represent the frames for learning the frame importance. However, hand-crafted features may not successfully represent semantic content of human-created summaries. Although some conventional supervised methods combine deep features of pre-trained deep learning models with different classifiers, task-specific high-level deep features for video summarization which can be learned by using deep learning networks are expected.

Recently, deep learning-based video summarization methods have been shown more effective to learn the human-created summaries. To model the correlations between continuous frames, the long short-term memory (LSTM)-based video summarization methods are proposed. By using

LSTM, the sequentially input frames are converted to frame importance scores as a sequence-to-sequence learning problem. For example, Zhang et al. [17] propose using the bidirectional LSTM (BiLSTM) architecture which is composed of forward and backward LSTMs to learn the continuous frame correlations. Zhao et al. [60] propose a hierarchical structure-adaptive RNN (HSA-RNN) by cascading two LSTM layers to achieve shot segmentation and video summarization. The first layer is developed for shot segmentation, while the second layer predicts the probability of each shot for video summarization. Wei et al. [61] propose a semantic attended video summarization network (SASUM) which contains a frame selector and LSTM-based encoder-decoder structure to translate the visual content to a text description and select semantic video segments. Ji et al. [21] propose attentive encoder–decoder networks by imposing a BiLSTM to encode features from input video frames and two attention-based LSTM to obtain the summary. Their method is improved in [20] by further considering the attentions of the LSTM features and a distribution consistent learning strategy.

The LSTM network is also combined with convolutional neural networks (CNNs) to extract spatial features for video summarization. For example, Zhou et al. [62] develop a deep summarization network which contains CNN and BiLSTM to learn video summaries based on a diversity-representativeness (DR) reward function computed from the generated summaries. They also provide supervised and unsupervised versions. Huang et al. [63] propose multi-stage spatiotemporal representations which combine 2D CNNs, 1D CNNs and LSTM to sequentially learn features to represent the frame-level importance. Modified from [63], Chu and Liu [64] propose a spatiotemporal modeling and label distribution (SMLD) learning method by adding the optical flow maps and user label distribution for frame representations. To model the long dependency among video frames, Wang et al. [65] propose a stacked memory network by using LSTM layers and memory layers to integrate the learned representations from CNNs.

Besides the LSTM-based methods, Yao et al. [16] propose a pairwise two-stream deep convolutional network to learn the relationships between high-light and non-highlight video segments. Rochan et al. [66] propose a fully convolutional sequence network (FCSN) to model the dependency among input frames and allow parallelization during processing frames. Both supervised and unsupervised versions are presented in [66]. Zhu et al. [18] propose a relational reasoning over spatial-temporal graphs (RR-STG) network to build spatial graphs on the detected object proposals and construct a temporal graph based on spatial graphs for importance score prediction of video frames.

More recently, attention-based video summarization methods are proposed and have been shown more effective for video summarization. Fajtl et al. [24] propose the VASNet which utilizes the self-attention mechanism to detect keyframes. Liu et al. [67] propose a hierarchical multi-attention network (H-MAN) which first applies BiLSTM to extract key-frame candidates and then applies a multi-attention model to generate the summary. In [68], a deep attentive and semantic preserving (DASP) method is proposed by combining BiLSTM with the attention module and restricting the semantic content of the summary to be consistent with the original video by using the mean square error. However, their attention mechanism only considers correlations between continuous frames. Informative positional embedding and multi-head self-attention schemes are not considered.

To consider temporal consistency for video summarization, Zhu et al. [19] propose a detect-to-summarize network (DSNet). Both anchor-based and anchor-free approaches are considered for object detection and used to predict the importance of video shots. Ghauri et al. [25] propose the multi-source visual attention (MSVA) scheme to integrate attention features of different types of inputs for video summarization. To model the frame dependency, Li et al. [22] consider pairwise temporal relations of video frames by using the attention scheme to generate diverse frames and form the video summary. Apostolidis et al. [26] encode the absolute position information with the self-attention mechanisms to represent frame dependency and assess the frame importance. Instead of considering CNN, LSTM or attention-based network, our preliminary work [31] proposes modifying the vision transformer [28] to achieve video summarization, which can better represent the correlations among frames. For more related papers and research issues about video summarization, please refer to the survey [52].

In summary, different from the attention-based methods which aim to learn the temporal frame dependency by using self-attentions, the proposed method imposes the frame embedding, index embedding and segment class embedding with the transformer-based TIA encoder to present the temporal inter-frame correlations among frames. Moreover, the proposed method collaborates the temporal correlations with the spatial intra-frame attention scheme by using the SIA encoder to further extract the salient regions which attract the human's attention from video frames. As a result, the proposed method can successfully outperform the state-of-the-art deep learning-based methods.

III. Spatiotemporal Vision Transformer

In this section, we will introduce the proposed spatiotemporal vision transformer (STVT). The framework of STVT is shown in Fig. 1. Given a training video, we divide the training video into non-overlapped video segments. Each segment serves as the input of the STVT to learn the human-created summaries. A pre-trained ResNet-18 network [69] is applied for each frame to extract deep frame features. These deep frame features are then linearly projected to 1D flattened frame features at first and then are projected by a trainable frame embedding projection to obtain the frame embedding. The frame embedding of the video segment is combined with the index embedding and segment class embedding to generate the embedded sequence which serves as the input of the TIA encoder. The TIA encoder aims to learn the temporal inter-frame correlations among frames in each segment by using multi-layer inter-frame attention modules. To further learn the spatial intra-frame attention, a SIA encoder is proposed with the spatial mapping module

and parallel intra-frame attention modules. The outputs of the intra-frame attention module are classified by using the frame importance classifiers. The learning of the network is driven by the proposed multi-frame loss in an end-to-end trainable manner. The weight-sharing of the parallel intra-frame attention modules and frame importance classifiers can help avoid overfitting and reduce the memory usage. Moreover, introducing fewer parameters of the SIA encoder also benefits the faster convergence during training. In the following, we will introduce each component in the STVT.

A. Embedded Sequence Module

Let each segment of the video contain N continuous frames. Let the image I contain N patches and the resolution of each patch is $W \times H$, where W and H are the frame width and height, respectively. The deep frame features of the nth frame are linearly projected to an 1D flattened frame feature $\boldsymbol{t}_j^n$ to represent the content of the nth frame in V_j. Then, $\boldsymbol{t}_j^n$ is mapped to a constant latent vector of size D with a trainable frame embedding projection $\boldsymbol{E}$. The output after the projection represents the frame embedding of the nth frame in V_j. To present the temporal information, the index embedding which represents the frame index correlations is applied to cooperate with the frame embedding. Finally, a segment class embedding is prepended to the frame embedding and boosts the global frame context information for learning the importance of the video segment.

To simultaneously represent the aforementioned information, the frame embedding, index embedding and segment class embedding of the video segment are combined to an embedded sequence z_0 as follows:

$$z_0 = [\boldsymbol{t}_j^{class}; \boldsymbol{t}_j^1\boldsymbol{E}; \boldsymbol{t}_j^2\boldsymbol{E}; \ldots; \boldsymbol{t}_j^N\boldsymbol{E}] + \boldsymbol{E}_{idx}, \tag{1}$$

where $\boldsymbol{t}_j^{class}$ is the segment class embedding and $\boldsymbol{t}_j^n$ is the 1D flattened frame feature of the nth frame of V_j. The dimension of $\boldsymbol{t}_j^n$ is $\mathcal{R}^{W\times H\times C}$. $\boldsymbol{E}$ is the trainable frame embedding projection, where the dimension of $\boldsymbol{E}$ is $\mathcal{R}^{(W\times H\times C)\times D}$. $\boldsymbol{E}_{idx}$ is the index embedding which contains the frame indices of frames in V_j. Because of the segment class embedding, the dimension of $\boldsymbol{E}_{idx}$ is $\mathcal{R}^{(N+1)\times D}$. To provide constant latent vector size and reduce dimension of the embedded sequence, $D = 768$ is set for the transformer layers in the TIA encoder as suggested in [28]. The embedded sequence z_0 serves as the input of the TIA encoder.

B. Temporal Inter-Frame Attention Encoder

The goal of the TIA encoder is to learn the temporal inter-frame correlations based on the multi-head self-attention of the transformer structure. The TIA encoder contains L multi-layer inter-frame attention modules. The inter-frame attention modules are modified from the transformer encoder [28] by replacing the patches of the image to frames of the segment. While the patch embedding in [28] aims to retain positional information, the index embedding in our approach aims to learn temporal correlations among frames. By using the idea, the importance of temporal frames can be effectively learned by the transformer encoder compared with recent attention-based methods. Moreover, the learned segment class embedding also provides global frame context representation for SIA encoders to extract intra-frame attentions.

Let $z_{\ell-1}$ be the input of the ℓth inter-frame attention module modified from the transformer encoder [28]. $\hat{z}_{\ell-1} = LN(z_{\ell-1})$ represents the normalization result of $z_{\ell-1}$, and $LN(\cdot)$ is the layer normalization function of the multi-head self-attention sub-module. Then, the multi-head self-attention of ℓth inter-frame attention module is defined as follows:

$$MSA(\hat{z}_{\ell-1}) = [SA_1(\hat{z}_{\ell-1}), \ldots, SA_M(\hat{z}_{\ell-1})]\boldsymbol{U}_t, \tag{2}$$

where SA_m is the mth self-attention head and $\boldsymbol{U}_t \in \mathcal{R}^{(M\cdot D_h)\times D}$ projects the concatenated feature to the multi-head attention. The frame importance can then be learned based on the temporal inter-frame correlations represented by the multi-head self-attention. As suggested in [27], we set $M = 12$ and $D_h = D/M$.

The output z'_ℓ of the multi-head self-attention sub-module is composed of the multi-head self-attention and a residual connection of the input as follows:

$$z'_\ell = MSA(\hat{z}_{\ell-1}) + z_{\ell-1}. \tag{3}$$

It serves as the input of the feed-forward feature sub-module. The feed-forward feature sub-module consists of a normalization layer, two fully connected layers and a residual connection. The output of the feed-forward feature sub-module is defined as:

$$z_\ell = MLP(\hat{z}'_\ell) + z'_\ell, \tag{4}$$

where $MLP(\cdot)$ represents the function of the fully connected layers, and $\hat{z}'_\ell = LN(z'_\ell)$. By concatenating L inter-frame attention modules, the temporal inter-frame correlations of the N frames of V_j can be learned by using the transformer-based TIA encoder. In this paper, we set $N = 16$ and $L = 12$ based on the evaluation results shown in Sec. IV-B. Finally, the output z_L of the last inter-frame attention module is sent to the SIA encoder for learning spatial intra-frame attention in the following.

C. Spatial Intra-Frame Attention Encoder

While the TIA encoder learns the temporal inter-frame correlations among frames, each frame may also contain salient regions which attract the human's attention. Such spatial attention within frames also helps represent the spatial information of human-created summaries to identify the frame importance. To learn the spatial intra-frame attention which represents spatial salient regions within each frame based on z_L, we propose the SIA encoder as shown in Fig. 2. Here, z_L represents the feature containing temporal correlations among frames with the segment class embedding information, so the dimension of z_L is $\mathcal{R}^{(N+1)\times D}$. Because we want to discover the spatial salient regions within each frame by using the SIA encoder, we need to decompose z_L to obtain individual features which can represent the spatial content with respect to each frame.

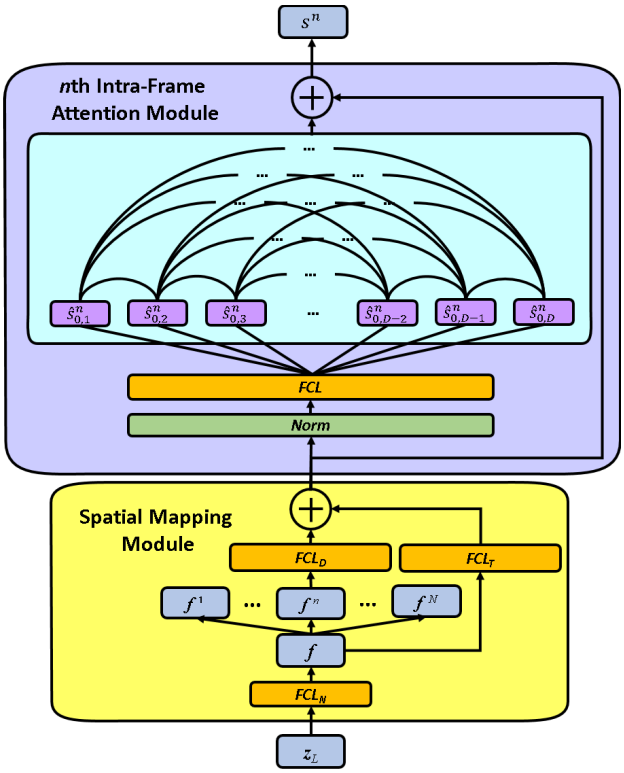


Fig. 2. The spatial intra-frame attention encoder includes the spatial mapping module and the parallel intra-frame attention modules. The spatial mapping module will map z_L to $\boldsymbol{f}$ at first. Here, we take $\boldsymbol{f}^n$ of the nth frame decomposed from $\boldsymbol{f}$ as an example to show how to obtain the frame transformer feature $\boldsymbol{s}_0^n$ of the nth frame. $\boldsymbol{s}_0^n$ serves as the input of the nth frame intra-frame attention module. The cyan rectangle represents the multi-head self-attention computed based on $\boldsymbol{s}_0^n$. Based on the multi-head self-attention, the spatial intra-frame attention within each frame can be effectively learned.

To obtain the feature of each frame of dimension $\mathcal{R}^{1\times D}$, a naive idea is to ignore the segment class embedding in z_L. In this way, a feature $\overline{z}_L$ of dimension $\mathcal{R}^{N\times D}$ can be obtained from z_L by removing the feature of the segment class embedding. Then, we can directly decompose the feature $\overline{z}_L$ to N features, where each decomposed feature of dimension $\mathcal{R}^{1\times D}$ only contains the frame embedding information. Nevertheless, the global frame context representation learned in the segment class embedding is not preserved and will be lost in the decomposed feature of each frame. Thus, the obtained spatial attention based on only the frame embedding of each frame may lead to discontinuous results between adjacent frames.

To solve the problem, a trainable spatial mapping module (SMM) is proposed to generate representative frame transformer features for each frame. The structure of the spatial mapping module is shown in Fig. 2. The first fully connected layer represented by a function $FCL_N(\cdot)$ contains N neurons to map z_L of dimension $\mathcal{R}^{(N+1)\times D}$ to a new feature $\boldsymbol{f}$ of dimension $\mathcal{R}^{N\times D}$. In this way, the segment class embedding information will also be combined to $\boldsymbol{f}$. $\boldsymbol{f}$ can be expressed as follows:

$$\boldsymbol{f} = [\boldsymbol{f}^1, \boldsymbol{f}^2, \ldots, \boldsymbol{f}^n, \ldots, \boldsymbol{f}^N], \tag{5}$$

where $\boldsymbol{f}^n$ represents the encoded feature of the nth frame of dimension $\mathcal{R}^{1\times D}$. Each encoded feature $\boldsymbol{f}^n$ is passed to a shared-weights fully connected layer $FCL_D(\cdot)$ of D neurons, where $FCL_D(\cdot)$ is the function representing the fully connected layer. To further integrate temporal continuous information of N frames, $\boldsymbol{f}$ is also passed to a fully connected layer of D neurons to obtain the temporal integration feature $FCL_T(\boldsymbol{f})$ of dimension $\mathcal{R}^{1\times D}$, where the function $FCL_T(\cdot)$ represents the function of the fully connected layer. Then, the frame transformer feature $\boldsymbol{s}_0^n$ of the nth frame is defined as follows:

$$\boldsymbol{s}_0^n = FCL_D(\boldsymbol{f}^n) + FCL_T(\boldsymbol{f}). \tag{6}$$

In this way, $\boldsymbol{s}_0^n$ contains the information of global frame context representation, continuous frame information and the frame content.

The frame transformer feature $\boldsymbol{s}_0^n$ of the nth frame is sent to the nth intra-frame attention module. The intra-frame attention module is also designed based on the transformer encoder to learn the spatial intra-frame attention for frame importance classification as shown in Fig. 2. The frame transformer feature is normalized by the normalization layer and $\hat{\boldsymbol{s}}_0^n = LN(\boldsymbol{s}_0^n)$ represents the layer normalization result of $\boldsymbol{s}_0^n$. Then, the multi-head self-attention of nth intra-frame attention module is defined as follows:

$$MSA(\hat{\boldsymbol{s}}_0^n) = [SA_1(\hat{\boldsymbol{s}}_0^n), \ldots, SA_M(\hat{\boldsymbol{s}}_0^n)]\boldsymbol{U}_s, \tag{7}$$

where SA_m is the mth self-attention head, $\boldsymbol{U}_s \in \mathcal{R}^{(M\cdot D_h)\times D}$ projects the concatenated feature to the multi-head attention, $M = 12$ and $D_h = D/M$.

The frame attention feature $\boldsymbol{s}^n$ of the nth frame is composed of the multi-head attention feature and a residual connection of the input feature as follows:

$$\boldsymbol{s}^n = MSA(\hat{\boldsymbol{s}}_0^n) + \boldsymbol{s}_0^n. \tag{8}$$

Based on the TIA encoder and SIA encoder, our method generates representative spatiotemporal frame features for frame importance classification.

D. Multi-Frame Loss

Given a training video segment V_j, we aim to learn the frame importance for each frame from human-created summaries. To individually learn the importance score of each frame, $\boldsymbol{s}^n$ serves as the feature of the frame importance classifier (FIC) which is composed of a fully connected layer of 2 neurons with a soft-max function for the binary frame importance classification. The frame loss $\mathcal{L}^n$ of the nth frame is defined as

$$\mathcal{L}^n = -\sum y^n \log_2(p^n), \tag{9}$$

where y^n is the ground truth frame importance, i.e. $y^n \in \{0, 1\}$ and p^n is the prediction result of the FIC of the nth frame, respectively.

To update the parameters of the proposed network based on the frame losses of the frames of the whole video segment, the multi-frame loss $\mathcal{L}$ is defined as the summarization of all frame losses as follows:

$$\mathcal{L} = \sum_{n=1}^{N} \mathcal{L}^n, \tag{10}$$

where N is the number of frames in each segment. Based on the multi-frame loss, our transformer-based network can be trained in an end-to-end trainable manner.

E. Video Summary Generation

During video summary generation, the frame importance is generated by using on the proposed spatiotemporal vision transformer. To produce the video summary of continuous content, we assess the scores of shots as suggested in [17] and [19]. To obtain shots, the kernel temporal segmentation (KTS) [3] is applied. The shot importance score c_i of the ith shot is defined as the average of the frame importance of the frames in the shot as follows:

$$c_i = \frac{1}{K_i} \sum_{k=1}^{K_i} p^k, \tag{11}$$

where p^k is the predicted frame importance of the kth frame, and K_i is the number of frames of the ith shot, respectively. To generate video summarization results of desired lengths for fair comparison, we follow [17] to set to the desired length as no more than 15% of the length of the original video. Then, the video summary generation problem can be considered as a {0/1} knapsack problem which can be solved by the following optimization process:

$$max \sum_{i=1}^{s_h} u_i c_i, \; s.t. \sum_{i=1}^{s_h} u_i K_i \leq V \times 15\% \tag{12}$$

where s_h is the number of shots, and V is the total length of the video. $u_i = 1$ means that the ith shot is selected to the video summary. Otherwise, $u_i = 0$ indicates that the ith shot is not important. The optimization can be solved by a dynamic programming approach and the final video summary is composed of the selected shots.

IV. Experimental Results

A. Datasets and Experimental Settings

1) Datasets: Two state-of-the-art video summarization datasets, SumMe [29] and TVSum [30], were used for evaluation. The SumMe dataset contains 25 videos and each video is annotated by at least 15 people. The TVSum dataset contains 50 videos of 10 categories and 1,000 annotations obtained by crowd-sourcing. Both datasets provide frame-level importance labels for evaluation. To provide fair comparisons, we followed the experimental dataset settings in [17] which were also applied in the state-of-the-art methods such as [18], [19], [20], and [21].

For the canonical experiments, the dataset was randomly divided into five splits. 80% of the dataset was used for training, the remaining 20% was used for testing. We performed our method five times and reported the average results as the dataset settings applied in [18], [19], [20], and [21]. For the augmented experiments, two additional training datasets OVP [70] and YouTube [70] were used. The OVP dataset contains 50 videos and the Youtube dataset contains 39 videos. Both datasets are annotated by 5 people with keyframe labels. Finally, for the transfer experiments, three datasets were used for training, while evaluating the remaining dataset. Following [17], we also apply the F-Score values for the performance evaluation and take the average of the testing videos which also follows the same dataset settings in [18], [19], [20], and [21].

2) Experimental Settings: We performed our experiments on an Intel i7 computer with a RTX 3090 GPU. Our method was implemented by using PyTorch 1.7.1. In our implementation, if the length of the last segment of the video is fewer than N, we repeat the last frame of the video until the length of the last segment of the video becomes N. In our method, a pre-trained ResNet-18 [69] on ImageNet [71] was used for feature extraction. All of the input frames were uniformly resized to 224×224. The number C of the deep frame features generated by a pre-trained ResNet-18 network is 512 for each frame. The batch size is 40 and the parameters of the proposed method are updated by using SGD with a momentum of 0.9. A dropout layer of dropout rate 0.1 is applied before the final classification layer. As suggested in [28], a warm-up mechanism and cosine learning rate are applied to avoid the overfitting during training the proposed network. The initial warm-up learning rate is set to 0.003. The learning is linearly increased to 0.03 at the 10th epoch. After the 10th epoch, the cosine learning rate decay is applied. Because no validation sets were provided in both datasets, we fixed the number of the training epochs to 100 without the selection of models.

B. Parameter Selection

The proposed method contains two adjustable parameters which will significantly affect the results. The first one is the number N of the input frames to the proposed network, where N also affects the number of the intra-frame attention modules of the SIA encoder. The second one is the number L of the inter-frame attention modules of the TIA encoder. In this subsection, we aim to evaluate the effects of N and L based on the SumMe and TVSum datasets.

Table I shows the results of different parameters. With the increasing number of the input frames N, the F-Score values also increase with respect to the same number L of the inter-frame attention modules of the TIA encoder. More input frames imply that the proposed network can learn more temporal inter-frame correlations among these frames. As a result, our method can achieve better results when N increases. By fixing N, with the increasing number of L, the F-Score values also increase. Such results show that more inter-frame attention modules help extract high-level semantic feature representations for temporal inter-frame correlations among frames during the frame importance learning. Based on the observations and the memory limitation of the hardware, we select $N = 16$ and $L = 12$ in the following experiments.

C. Ablation Study

In our method, the SIA encoder contains the spatial mapping module (SMM) to learn the frame transformer feature and the intra-frame attention (IFA) module to learn the spatial intra-frame attention for frame importance classification. Moreover, the dynamic learning rates (DLR) are applied during training to avoid overfitting. In the ablation study, we aim to evaluate

TABLE I
COMPARISONS OF DIFFERENT PARAMETERS IN F-SCORE (%)

N	L	SumMe	TVSum
9	4	47.1	60.4
	8	48.9	63.6
	12	51.8	64.8
16	4	51.6	61.9
	8	52.2	64.6
	12	**55.1**	**67.1**

TABLE II
ABLATION STUDY OF PROPOSED SCHEMES IN F-SCORE (%)

SMM	IFA	DLR	SumMe	TVSum
✓			50.0	60.2
	✓		50.5	60.8
		✓	48.1	59.4
✓	✓		53.2	65.8
✓		✓	52.2	65.0
	✓	✓	52.5	65.5
✓*	✓	✓	51.5	65.9
✓	✓	✓	**55.1**	**67.1**

TABLE III
ABLATION STUDY OF DIFFERENT BACKBONES AND DOWN-SAMPLING IN F-SCORE (%)

Backbone	Down-sampling	SumMe	TVSum
ResNet-18	Yes	51.2	65.0
ResNet-18	No	**55.1**	**67.1**
GoogLeNet	No	53.7	66.3

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART METHODS IN F-SCORE (%)

Methods	SumMe		TVSum		AveRank
	F1	Rank	F1	Rank	
LiveLight [11]	-	-	46.0	27	27
MSDS-CC [40]	40.6	25	52.3	26	25.5
DySeqDpp [55]	44.3	18	58.4	18	18
Zhang et al. [57]	40.9	24	-	-	24
Li et al. [58]	43.1	21	52.7	25	23
HSA-RNN [60]	44.1	19	59.8	15	17
vsLSTM [17]	37.6	27	54.2	24	25.5
dppLSTM [17]	38.6	26	54.7	23	24.5
SUM-GAN [45]	41.7	23	56.3	22	22.5
DR-DSN [62]	42.1	22	58.1	20	21
SMLD [64]	47.6	11	61.0	11	11
SASUM [61]	45.3	16	58.2	19	17.5
FCSN [66]	47.5	12	56.8	21	16.5
A-AVS [21]	43.9	20	59.4	16	18
M-AVS [21]	44.4	17	61.0	11	14
ADSum-A [20]	45.9	14	64.5	2	8
ADSum-M [20]	46.1	13	64.3	3	8
DSNet (Based) [19]	50.2	8	62.1	7	7.5
DSNet (Free) [19]	51.2	7	61.9	8	7.5
H-MAN [67]	51.8	6	60.4	14	10
DASP [68]	45.5	15	63.6	4	9.5
VASNet [24]	49.7	9	61.4	10	9.5
MSVA [25]	53.4	3	61.5	9	6
SUM-GDA [22]	52.8	5	58.9	17	11
PGL-SUM [26]	**55.6**	**1**	61.0	11	6
RR-STG [18]	53.4	3	63.0	5	4
Preliminary [31]	49.0	10	62.3	6	8
Proposed	55.1	2	**67.1**	**1**	**1.5**

the effectiveness of SMM, IFA and DLR. Table II shows the results of the ablation study. The first, second and third rows show the F-Score values of only using SMM, IFA and DLR, respectively, for the SumMe and TVSum datasets. When only considering SMM or IFA in the proposed network, the results are better than that of only considering DLR during training. These results imply the importance and effectiveness of the design of the SMM and IFA modules in the proposed SIA decoder.

When combining SMM and IFA schemes, the results are better than those of remaining two combinations. These results are consistent with the aforementioned results. The results of the proposed method without considering the segment class embedding in SMM are shown in the seventh row of Table II. Because the segment class embedding contains global frame context information, it can also help the learning of the spatiotemporal features to represent frame importance. Thus, the performance of the proposed method without considering the segment class embedding then drops. By simultaneously using three schemes and the segment class embedding, the proposed method achieves the best F-Score values in both datasets.

Table III shows the ablation study of the proposed method by using the down-sampling scheme and different backbones. The results of the proposed method with down-sampling is shown in the first row of Table III. When applying down-sampling to training frames, the number of training frames will be significantly reduced. Important temporal inter-frame and spatial intra-frame information may not be learned by the proposed transformer. Thus, the performance of the proposed method with down-sampling is significantly lower than that of the proposed method without down-sampling as shown in the second row of Table III. Such results show that the proposed method can be benefited from the full set of the video frames.

To further investigate the effects of different CNN backbones used in the embedded sequence module, we apply GoogLeNet [72] pre-trained in ImageNet [71] as the feature backbone which is commonly applied in recent video summarization methods [19], [20], [21]. As shown in Table III, when applying ResNet-18 as the feature backbone, the proposed method can achieve better results. Nevertheless, the performance gap is small which implies that the proposed transformer can learn the representative spatial and temporal information from both of the CNN features.

D. Quantitative Results

1) Comparisons: To show the effectiveness of the proposed method, we compared our method with the state-of-the-art methods in both SumMe and TVSum datasets. Specifically, we compared the proposed method with three conventional methods including LiveLight [11], MSDS-CC [40] and DySeqDpp [55]. Because deep learning-based video summarization methods have been shown better performance compared with conventional methods, we also compared our method with deep learning-based methods of different backbone networks. In the comparisons, the state-of-the-art deep learning methods including Zhang et al. [57], Li et al. [58], HSA-RNN [60], vsLSTM [17], dppLSTM [17], SUM-GAN [45], DR-DSN [62], SMLD [64], SASUM [61], FCSN [66], A-AVS [21], M-AVS [21], ADSum-A [20], ADSum-M [20], anchor-based DSNet [19], anchor-free

DSNet [19], H-MAN [67], DASP [68], VASNet [24], MSVA [25], SUM-GDA [22], PGL-SUM [26], RR-STG [18], and the preliminary method [31] are compared. The results of each competing method are shown in Table IV. For fair comparisons, all of the results of the competing methods listed in the section are obtained from the original papers of the competing methods.

As shown in Table IV, the proposed method outperforms these state-of-the-art methods in both SumMe and TVSum datasets by considering the average rank (AveRank) [52] on both datasets. Compared with LSTM- and attention-based methods which model temporal information of frames, the proposed method can better learn the temporal inter-frame correlations and spatial intra-frame attention based on the novel spatiotemporal vision transformer for frame importance classification. Moreover, combining the TIA encoder with the SIA encoder significantly improves the performance of our preliminary method. The results also reveal the effectiveness of the transformer structure for solving the video summarization problem compared with the LSTM- and attention-based methods. Thus, our method can effectively learn the frame importance from human-created summaries and achieve the best video summarization results in both datasets.

Besides the F-Score values, Kendall's τ and Spearman's ρ correlation coefficients are applied in [73] to assess the similarities between the ranks provided by the human annotated and model generated frame importance scores. Table V shows the quantitative results of the state-of-the-art methods WS-HRL [51], RSGN [74], DAC [75], DAN [76], HMT [77] and the proposed method with respect to Kendall's τ and Spearman's ρ correlation coefficients in the TVSum dataset. As shown in [51], [73], and [75], these correlation coefficients are computed based on the ranking results of predicted video summarization from uniformly sampled videos. To simulate the predicted video summarization results as the settings in [51], [73], and [75] for fair comparisons, we uniformly sampled and ranked our video summarization results based on predicted frame importance and human annotated frame importance. The final correlation coefficients are obtained by averaging the ranking results of individual human subjects. Because the proposed method can learn the inter-frame and intra-frame attentions by using the TIA encoder and SIA encoder, the predicted frame importance scores are consistent with the human perception. Therefore, the ranks provided by the proposed method can achieve better results compared with the competing methods.

The average training time of the proposed method with the full set of video frames for the SumMe and TVSum datasets are 1 hour and 3 hours, respectively. The average inference speed of the proposed method with the full set of video frames for the SumMe and TVSum datasets are 329 and 325 frames per-second (FPS), respectively. Based on the quantitative results, we show that the proposed method can achieve better performance by learning the full set of the videos and real-time processing efficiency during testing. In addition, the number of parameters of the proposed STVT is 90M which is only slightly larger than that of the vision transformer (86M). Thus, our STVT can be run on general GPU devices.

TABLE V
COMPARISONS WITH STATE-OF-THE-ART METHODS FOR KENDALL'S AND SPEARMAN'S CORRELATION COEFFICIENTS

Methods	Kendall's τ	Spearman's ρ
dppLSTM [17]	0.042	0.055
DR-DSN [62]	0.020	0.026
WS-HRL [51]	0.078	0.116
RSGN [74]	0.083	0.090
DAC [75]	0.058	0.065
DAN [76]	0.071	0.099
HMT [77]	0.096	0.107
Proposed	**0.100**	**0.131**

TABLE VI
COMPARISONS WITH STATE-OF-THE-ART METHODS FOR CANONICAL (C), AUGMENTED (A) AND TRANSFER (T) SETTINGS IN F-SCORE (%)

Methods	SumMe			TVSum		
	C	A	T	C	A	T
vsLSTM [17]	37.6	41.6	40.7	54.2	57.9	56.9
dppLSTM [17]	38.6	42.9	41.8	54.7	59.6	58.7
SUM-GAN [45]	41.7	43.6	–	56.3	61.2	–
DR-DSN [62]	42.1	43.9	42.6	58.1	59.8	58.9
FCSN [66]	47.5	51.1	44.1	56.8	59.2	58.2
A-AVS [21]	43.9	44.6	–	59.4	60.8	–
M-AVS [21]	44.4	46.1	–	61.0	61.8	–
ADSum-A [20]	45.9	47.3	–	64.5	65.8	–
ADSum-M [20]	46.1	47.6	–	64.3	65.7	–
DSNet (Based) [19]	50.2	50.7	46.5	62.1	63.9	59.4
DSNet (Free) [19]	51.2	53.3	47.6	61.9	62.2	58.0
H-MAN [67]	51.8	52.5	48.1	60.4	61.0	59.5
SUM-GDA [22]	52.8	54.4	46.9	58.9	60.1	59.0
DASP [68]	45.5	47.0	-	63.6	64.5	-
RR-STG [18]	53.4	54.8	45.4	63.0	63.6	59.7
Proposed	**55.1**	**55.9**	**48.2**	**67.1**	**67.7**	**59.9**

2) Augmented and Transfer Experiments: To further evaluate the performance of the proposed method, we performed experiments on augmented and transfer datasets by following the augmented and transfer settings in [19] and [20]. Using augmented data can help reduce the influence of insufficient training data and the overfitting problem. As described in [18], [19], [20], and [21], two additional datasets OVP and YouTube datasets [70] are served as the augmented datasets with the SumMe and TVSum datasets for augmented experiments.

Because not all of the aforementioned competing methods perform the augmented experiments, we only list the competing methods which also perform the augmented experiments. With the augmented datasets, the performance of all of the competing methods can be improved as shown in Table VI. Such results show that more training videos help the learning of video summarization. By using the augmented data for training, the performance of the proposed method improves 0.8% and 0.6% for the SumMe and the TVSum datasets, respectively. Moreover, the proposed method consistently achieves the best performance compared with the state-of-the-art methods in the augmented experiments.

To provide more challenging experiments, we performed the experiments based on the transfer settings. As shown in Table VI, the proposed method still outperforms the state-of-the-art methods. The results show the effectiveness of the proposed method to classify frame importance across different datasets.

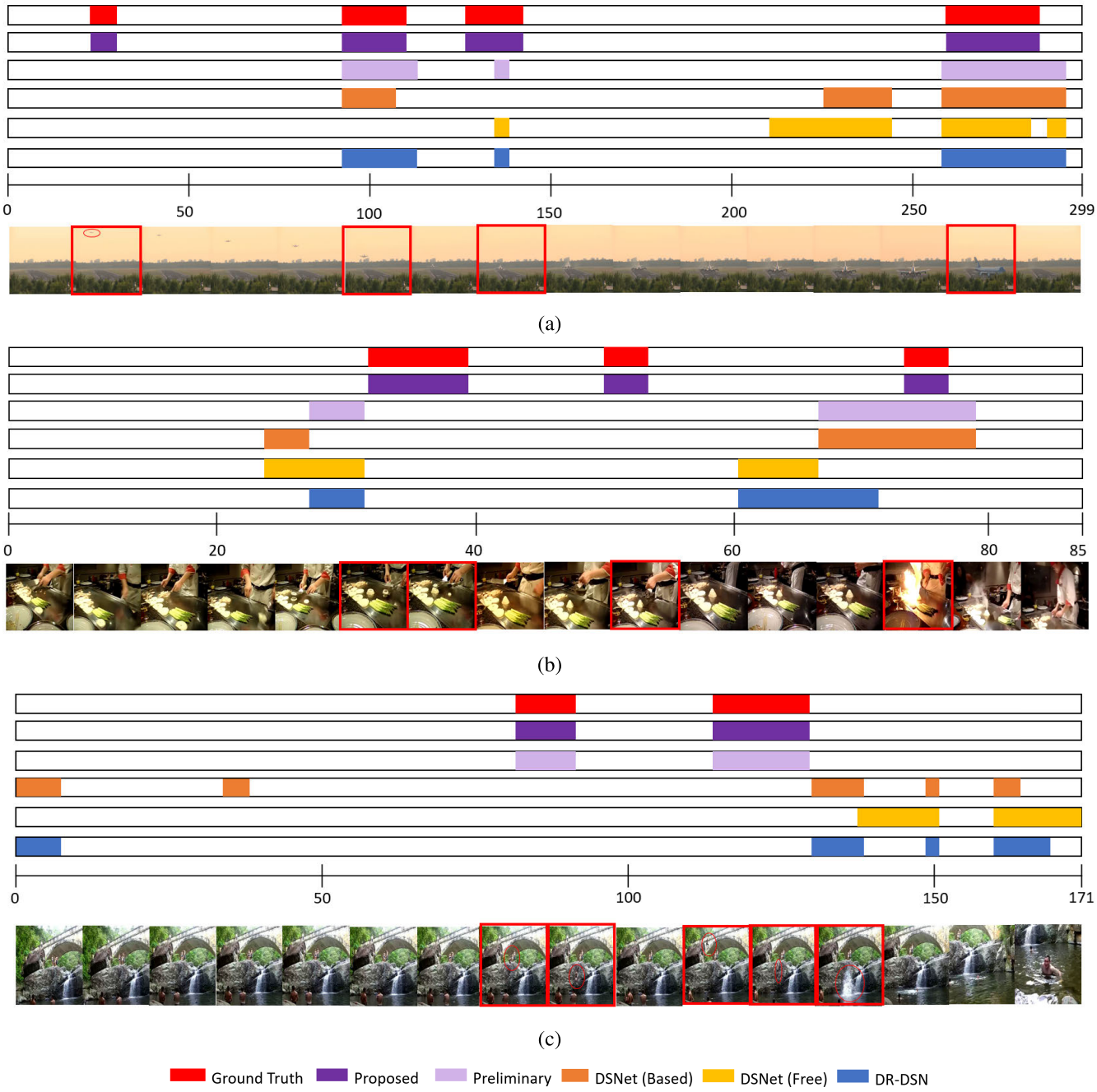


Fig. 3. The qualitative results of competing methods for video summarization including the ground truth (Red), the proposed method (Purple), the preliminary method (Light Purple) [31], anchor-based DSNet (Orange) [19], anchor-free DSNet (Yellow) [19] and DR-DSN (Blue) [62]. The x-axis of the figure represents the frame indices. The qualitative results of (a) the 1st video, (b) the 8th video, and (c) the 24th video of the SumMe dataset are presented, respectively.

E. Qualitative Results

To show frame based summarization results, we compared the proposed method with the preliminary method [31], anchor-based DSNet [19], anchor-free DSNet [19] and DR-DSN [62] which provided the source codes for generating summarization results. Fig. 3(a), (b) and (c) show the qualitative results of the 1st, the 8th and the 24th videos of the SumMe dataset, respectively. As shown in Fig. 3(a), an airplane is landing. When the airplane firstly appears in the video, it attracts the human's attention as labelled by the ground truth. Due to the smaller spatial size of the airplane, the competing methods fail to summarize this event. Because of the SIA encoder, the proposed method can extract intra-frame attention to successfully summarize this event. Moreover, the TIA encoder can learn the temporal inter-frame correlations from the full set of video frames. Thus, the lengths of the summarized events can be more consistent with those of the ground truth. Similar situations can also be observed in Fig. 3(b). The proposed method successfully summarizes the events of the teppanyaki chef compared with the competing methods. As shown in Fig. 3(c), the ground truth summaries the frames when people jump to the river. The generated video

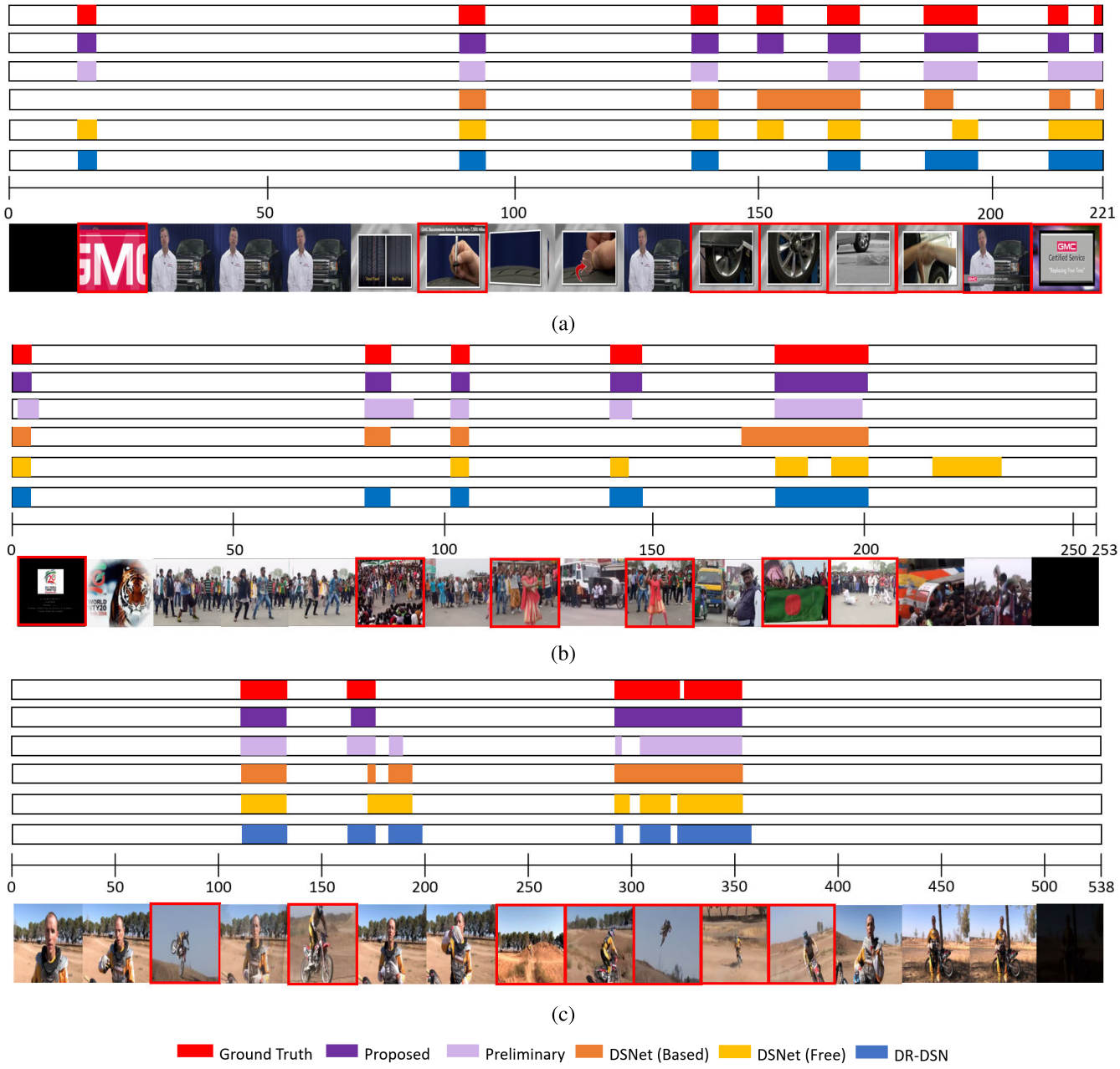


Fig. 4. The qualitative results of competing methods for video summarization including the ground truth (Red), the proposed method (Purple), the preliminary method (Light Purple) [31], anchor-based DSNet (Orange) [19], anchor-free DSNet (Yellow) [19] and DR-DSN (Blue) [62]. The x-axis of the figure represents the frame indices. The qualitative results of (a) the 5th video, (b) the 32th video, and (c) the 41th video of the TVSum dataset are presented, respectively.

summarization results of the proposed method are again more consistent with the ground truth summaries.

Fig. 4(a), (b) and (c) show the qualitative results of the 5th, the 32th and 41th videos of the TVSum dataset, respectively. Compared with the competing methods, the proposed method achieves the most consistent results with the ground truth in both of the frame indices and the lengths of the summaries. These qualitative results show the effectiveness of the transformer-based TIA encoder and SIA encoder in the proposed method compared with the LSTM-based methods.

Fig. 5 shows a false summarization result of the proposed method. In the 19th video of the SumMe dataset, people label important frames when the airplane flies near the cameraman during landing. These important frames are correctly captured by the proposed method. When the airplane first appears in the video (before the 50th frame), the false summarization occurs. Because the appearance of the airplane is significantly different from those of the sky and sea, the airplane can be considered as a salient object either from temporal views between adjacent frames or spatial views within the frame content. Thus, the salient regions of the frames containing the airplane are captured by the proposed spatiotemporal self-attention scheme and lead to the false summarization. Nevertheless, the proposed method still outperforms the competing methods as shown in Fig. 5.

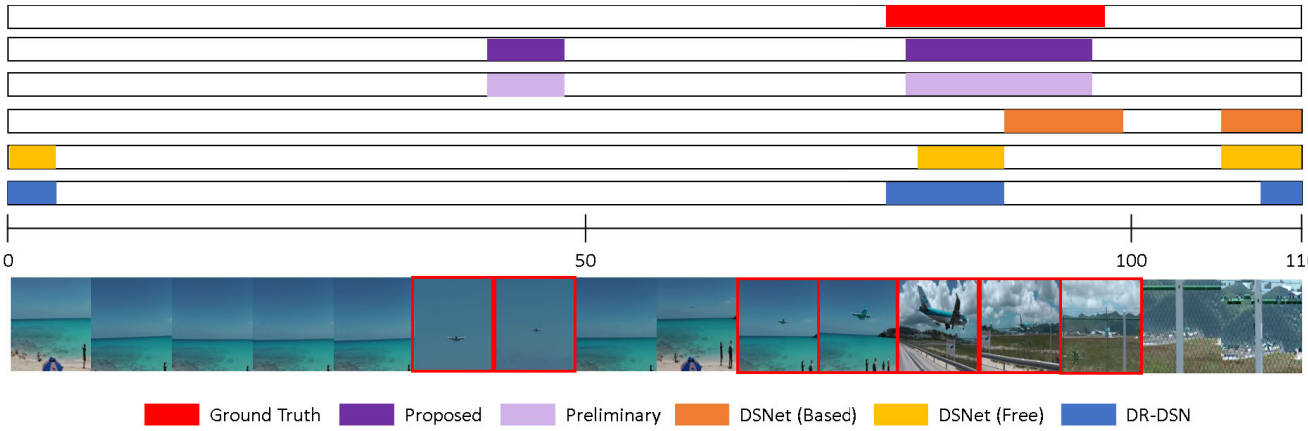
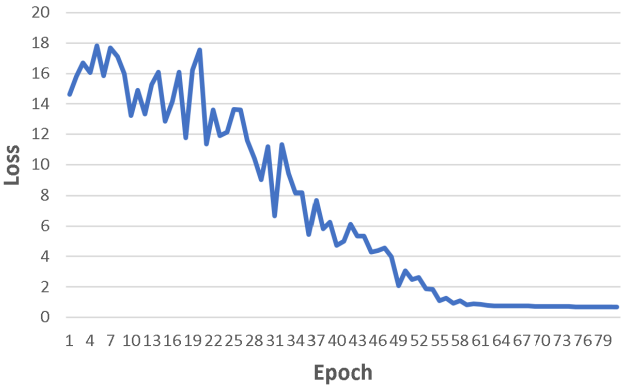
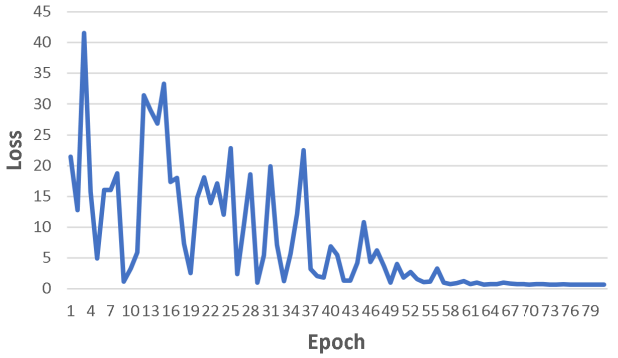


Fig. 5. The qualitative results of competing methods for video summarization including the ground truth (Red), the proposed method (Purple), the preliminary method (Light Purple) [31], anchor-based DSNet (Orange) [19], anchor-free DSNet (Yellow) [19] and DR-DSN (Blue) [62]. The x-axis of the figure represents the frame indices. The qualitative results of the 19th video of the SumMe dataset.



(a)



(b)

Fig. 6. The loss plots of (a) training and (b) validation.

F. Discussion

Fig. 6 shows the plots of the training loss and validation loss. Please note that in the proposed method, we train the network with the fixed number of epochs without using model selection strategies to select models with better results. With the proposed shared-weights, warm-up mechanism and cosine learning rate, the training loss can gradually decrease as shown in Fig. 6(a) and converge when the network is trained with sufficient epochs. Similar situations can also be observed in Fig. 6(b) for validation. In addition, because we use the full set of the video frames for training, the proposed network can fully utilize the frames to learn the spatiotemporal correlations with respect to human-created summaries. Thus, the overfitting problem can also be alleviated. Such results can also be observed from Table III. As a result, the proposed method achieves the best results by training from the full set of the training videos.

V. Conclusion

In this paper, we propose a novel spatiotemporal vision transformer which consists of the embedded sequence module, TIA encoder and SIA encoder for video summarization. While the TIA encoder learns the temporal inter-frame correlations among adjacent and non-adjacent frames, the SIA encoder extracts spatial intra-frame attention of each frame to represent the frame importance. By composing these two encoders, the proposed method can learn more representative spatiotemporal features from human-created summaries for video summarization. As shown in the experiments, the proposed transformer-based network outperforms competing deep learning-based methods in both of the SumMe and TVSum datasets. The results also reveal the effectiveness of using novel transformer-based structure for video summarization. In the future, we will apply the proposed method to solve video content analysis problems such as video saliency analysis and video synopsis.

References

[1] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. Int. Conf. Image Process.*, 1998, pp. 866–870.

[2] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 66–75, Feb. 2012.

[3] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 540–555.

[4] R. Cong, J. Lei, H. Fu, M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[5] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1336–1349, Aug. 2014.

[6] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1971–1984, Nov. 2008.

[7] C.-R. Huang, P.-C.-J. Chung, D.-K. Yang, H.-C. Chen, and G.-J. Huang, "Maximum a posteriori probability estimation for online surveillance video synopsis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1417–1429, Aug. 2014.

[8] S. Xiao, Z. Zhao, Z. Zhang, Z. Guan, and D. Cai, "Query-biased self-attentive network for query-focused video summarization," *IEEE Trans. Image Process.*, vol. 29, pp. 5889–5899, 2020.

[9] Y. Hu, M. Liu, X. Su, Z. Gao, and L. Nie, "Video moment localization via deep cross-modal hashing," *IEEE Trans. Image Process.*, vol. 30, pp. 4667–4677, 2021.

[10] A. Aner and J. R. Kender, "Video summaries through mosaic-based shot and scene clustering," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 388–402.

[11] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2513–2520.

[12] Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon, "Discriminative feature learning for unsupervised video summarization," in *Proc. Nat. Conf. Artif. Intell.*, 2019, pp. 1–22.

[13] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3584–3592.

[14] R. Panda, A. Das, Z. Wu, J. Ernst, and A. K. Roy-Chowdhury, "Weakly supervised summarization of web videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3677–3686.

[15] S. Cai, W. Zuo, L. S. Davis, and L. Zhang, "Weakly supervised video summarization using variational encoder–decoder and web prior," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 184–200.

[16] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 982–990.

[17] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 766–782.

[18] W. Zhu, Y. Han, J. Lu, and J. Zhou, "Relational reasoning over spatial–temporal graphs for video summarization," *IEEE Trans. Image Process.*, vol. 31, pp. 3017–3031, 2022.

[19] W. Zhu, J. Lu, J. Li, and J. Zhou, "DSNet: A flexible detect-to-summarize network for video summarization," *IEEE Trans. Image Process.*, vol. 30, pp. 948–962, 2021.

[20] Z. Ji, Y. Zhao, Y. Pang, X. Li, and J. Han, "Deep attentive video summarization with distribution consistency learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1765–1775, Apr. 2021.

[21] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder–decoder networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1709–1717, Jun. 2020.

[22] P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, "Exploring global diverse attention via pairwise temporal relation for video summarization," *Pattern Recognit.*, vol. 111, Mar. 2021, Art. no. 107677.

[23] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5457–5466.

[24] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Proc. Asian Conf. Comput. Vis. Workshops*, 2019, pp. 39–54.

[25] J. A. Ghauri, S. Hakimov, and R. Ewerth, "Supervised video summarization via multiple feature sets with parallel attention," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[26] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, "Combining global and local attention with positional encoding for video summarization," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Nov. 2021, pp. 226–234.

[27] V. Ashish et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[28] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–14.

[29] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 505–520.

[30] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5179–5187.

[31] T.-C. Hsu, Y.-S. Liao, and C.-R. Huang, "Video summarization with frame index vision transformer," in *Proc. 17th Int. Conf. Mach. Vis. Appl. (MVA)*, Jul. 2021, pp. 1–5.

[32] W. Wolf, "Key frame selection by motion analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, Mar. 1996, pp. 1228–1231.

[33] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using Delaunay clustering," *Int. J. Digit. Libraries*, vol. 6, no. 2, pp. 219–232, Apr. 2006.

[34] X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 9, pp. 765–776, Sep. 2002.

[35] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 975–985, Aug. 2012.

[36] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang, "Automatic video summarization by graph modeling," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 104–109.

[37] Y. Hadi, F. Essannouni, and R. O. H. Thami, "Video summarization by k-medoid clustering," in *Proc. ACM Symp. Appl. Comput.*, Apr. 2006, pp. 1400–1401.

[38] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1600–1607.

[39] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng, "Video summarization via minimum sparse reconstruction," *Pattern Recognit.*, vol. 48, no. 2, pp. 522–533, Feb. 2015.

[40] J. Meng, S. Wang, H. Wang, Y.-P. Tan, and J. Yuan, "Video summarization via multi-view representative selection," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1189–1198.

[41] R. Panda and A. K. Roy-Chowdhury, "Sparse modeling for topic-oriented video summarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1388–1392.

[42] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. 10th ACM Int. Conf. Multimedia*, Dec. 2002, pp. 533–542.

[43] N. Ejaz, I. Mehmood, and S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Process., Image Commun.*, vol. 28, no. 1, pp. 34–44, Jan. 2013.

[44] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–13.

[45] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2982–2991.

[46] L. Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, "Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization," in *Proc. Nat. Conf. Artificial Intell.*, Jul. 2019, vol. 33, no. 1, pp. 9143–9150.

[47] Y. Jung, D. Cho, S. Woo, and I. S. Kweon, "Global-and-local relative position embedding for unsupervised video summarization," in *Proc. Eur. Conf. Comput. Vis.*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., 2020, pp. 167–183.

[48] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3278–3292, Aug. 2021.

[49] Y. Zhang, X. Liang, D. Zhang, M. Tan, and E. P. Xing, "Unsupervised object-level video summarization with online motion auto-encoder," *Pattern Recognit. Lett.*, vol. 130, pp. 376–385, Feb. 2020.

[50] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using web-image priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2698–2705.

[51] Y. Chen, L. Tao, X. Wang, and T. Yamasaki, "Weakly supervised video summarization by hierarchical reinforcement learning," in *Proc. ACM Multimedia Asia*, New York, NY, USA, 2020, pp. 1–6.

[52] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *Proc. IEEE*, vol. 109, no. 11, pp. 1838–1863, Nov. 2021.

[53] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3090–3098.

[54] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 2014, pp. 2069–2077.
[55] Y. Li, L. Wang, T. Yang, and B. Gong, "How local is the local diversity? Reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 156–174. [Online]. Available: https://dblp.org/rec/conf/eccv/LiWYG18.html
[56] A. Sharghi, A. Borji, C. Li, T. Yang, and B. Gong, "Improving sequential determinantal point processes for supervised video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 517–533.
[57] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1059–1067.
[58] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3652–3664, Aug. 2017.
[59] A. Sharghi, B. Gong, and M. Shah, "Query-focused extractive video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 3–19.
[60] B. Zhao, X. Li, and X. Lu, "HSA-RNN: Hierarchical structure-adaptive RNN for video summarization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7405–7414.
[61] H. Wei, B. Ni, Y. Yan, H. Yu, and X. Yang, "Video summarization via semantic attended networks," in *Proc. Nat. Conf. Artif. Intell.*, 2018, pp. 216–223.
[62] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. Nat. Conf. Artif. Intell.*, 2018, pp. 7582–7589.
[63] S. Huang, X. Li, Z. Zhang, F. Wu, and J. Han, "User-ranking video summarization with multi-stage spatio-temporal representation," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2654–2664, Jun. 2019.
[64] W. Chu and Y. Liu, "Spatiotemporal modeling and label distribution learning for video summarization," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2019, pp. 1–6.
[65] J. Wang, W. Wang, Z. Wang, L. Wang, D. Feng, and T. Tan, "Stacked memory network for video summarization," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, p. 836.
[66] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 358–374.
[67] Y. Liu, Y. Li, F. Yang, S. Chen, and Y. F. Wang, "Learning hierarchical self-attention for video summarization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3377–3381.
[68] Z. Ji, F. Jiao, Y. Pang, and L. Shao, "Deep attentive and semantic preserving video summarization," *Neurocomputing*, vol. 405, pp. 200–207, Sep. 2020.
[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[70] S. E. F. de Avila, A. P. B. Lopes, A. da Luz, and A. de Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 56–68, Jan. 2011.
[71] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[72] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[73] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkilä, "Rethinking the evaluation of video summaries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7588–7596.
[74] B. Zhao, H. Li, X. Lu, and X. Li, "Reconstructive sequence-graph network for video summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2793–2801, May 2022.
[75] H. Fu, H. Wang, and J. Yang, "Video summarization with a dual attention capsule network," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 446–451.
[76] G. Liang, Y. Lv, S. Li, X. Wang, and Y. Zhang, "Video summarization with a dual-path attentive network," *Neurocomputing*, vol. 467, pp. 1–9, Jan. 2022.
[77] B. Zhao, M. Gong, and X. Li, "Hierarchical multimodal transformer to summarize videos," *Neurocomputing*, vol. 468, pp. 360–369, Jan. 2022.

Tzu-Chun Hsu received the B.S. degree from the Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan, in 2020, and the M.S. degree from the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, in 2022.

Yi-Sheng Liao received the B.S. and M.S. degrees from the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan, in 2018 and 2020, respectively.

Chun-Rong Huang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 1999 and 2005, respectively. In 2005, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, as a Postdoctoral Fellow. He joined the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan, in 2010, where he became a Full Professor in 2019. In 2023, he joined the Cross College Elite Program, Academy of Innovative Semiconductor and Sustainable Manufacturing, National Cheng Kung University, Tainan. His research interests include computer vision, computer graphics, multimedia signal processing, image processing, and medical image processing. He is a member of the IEEE Circuits and Systems Society, the IEEE Signal Processing Society, the IEEE Computational Intelligence Society, and the Phi Tau Phi Honor Society.