Semantic representation and attention alignment for Graph Information Bottleneck in video summarization

Rui Zhong, *Member, IEEE,* Rui Wang, Wenjin Yao, Min Hu, *Student Members, IEEE,* Shi Dong, Adrian Munteanu

***Abstract*—End-to-end Long Short-Term Memory (LSTM) has been successfully applied to video summarization. However, the weakness of the LSTM model, poor generalization with inefficient representation learning for inputted nodes, limits its capability to efficiently carry out node classification within user-created videos. Given the power of Graph Neural Networks (GNNs) in representation learning, we adopted the Graph Information Bottle (GIB) to develop a Contextual Feature Transformation (CFT) mechanism that refines the temporal dual-feature, yielding a semantic representation with attention alignment. Furthermore, a novel Salient-Area-Size-based spatial attention model is presented to extract frame-wise visual features based on the observation that humans tend to focus on sizable and moving objects. Lastly, semantic representation is embedded within attention alignment under the end-to-end LSTM framework to differentiate indistinguishable images. Extensive experiments demonstrate that the proposed method outperforms State-Of-The-Art (SOTA) methods.**

***Index Terms*—Graph Information Bottleneck, Contextual Feature Transformation (CFT), Spatial Attention Model, video summarization, Bi-LSTM.**

I. Introduction

In recent years, an enormous amount of user-created videos have been spread widely online through major social media platforms, including YouTube, Instagram, and Tiktok. Social media platforms accelerate the trend of browsing short videos with a wide coverage of video topics. In Fig. 1 (a), we illustrate a common characteristic of user-created videos, that is, that it contains much richer content than the surveillance video. Prior works have made attempts to efficiently perform video summarization [1], [2], [3]. However, most existing approaches are not efficient enough to deal with user-created videos. Therefore, we propose a Semantic representation and Attention alignment-based Contextual Feature Transformation (SA-CFT) with the Graph Information Bottleneck (GIB) model in video summarization.

The classic video summarization techniques have applied machine learning to develop models that capture intricate patterns of video cues, such as visual attention [4], foreground objects, and motion cues [5], [6]. One of the mainstream methods treats video summarization as a clustering problem, where the frames containing content closer to the clustering center are selected as the summary [7]. The classic methods' efficiency relied on the representation ability of hand-crafted feature extractors. However, hand-crafted feature extractors could only be designed for small datasets and particular video domains, such as sports games, news, movies, and surveillance videos. They must be cautiously designed with appropriate expertise. Due to the weakness of their representation ability, classic methods have their limitations when dealing with large datasets with more complicated content.

Researchers have proposed deep neural network-based summarization methods to overcome the disadvantages of classic methods. Instead of being constrained to particular video domains, deep neural network-based methods could be adopted for general video by establishing objective assessment factors, such as diversity, representativeness, continuity, and so on.

Currently, video summarization methods based on deep neural network models have roughly been categorized into supervised and unsupervised methods. The supervised video summarization methods leverage deep neural networks to process the inputted video under the guidance of human-annotated datasets during training [8], [9], [10]. When high-quality annotated data sets are insufficient, the supervised methods may suffer from poor performance. However, producing high-quality annotated datasets is highly labor-intensive.

In unsupervised methods, the hand-crafted criteria, such as diversity and representativeness, have been designed to replace the ground-truth labels. While video summaries are evaluated in terms of their similarity with human-annotated summaries, the hand-crafted criteria should capture the essential mechanism of human annotation in selecting keyframes. To promote these unsupervised methods, Zhou presented the reinforcement learning-based end-to-end LSTM method [11] as an effective way to generate summaries. The key idea was to design a human-crafted reward similar to the mechanism of manual annotation and gradually enhance the summarization performance during reinforcement learning. Due to the difficulty of designing a suitable reward mechanism, reinforcement learning-based methods are facing difficulties [12].

However, the baseline network of Zhou's method [11], the end-to-end LSTM, has efficiently preserved the advantages of LSTM in capturing long-range dependencies and further

Rui Zhong, Rui Wang, Wenjin Yao, Shi Dong are with Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, School of Computer Science, Central China Normal University, Wuhan, 430079, Hubei, China (e-mail: zhongrui0824@126.com, wangrui9891@foxmail.com, yaowj@ccnu.edu.cn, dongshi@ccnu.edu.cn).

Min Hu is with National Engineering Research Center for Multimedia Software, Wuhan University, China. (e-mail: humin0328@163.com).

Adrian Munteanu is with imec-VUB, Electronics and Informatics (ETRO) Department, Vrije Universiteit Brussel (VUB), 1050 Ixelles, Belgium (e-mail: Adrian.Munteanu@vub.be).

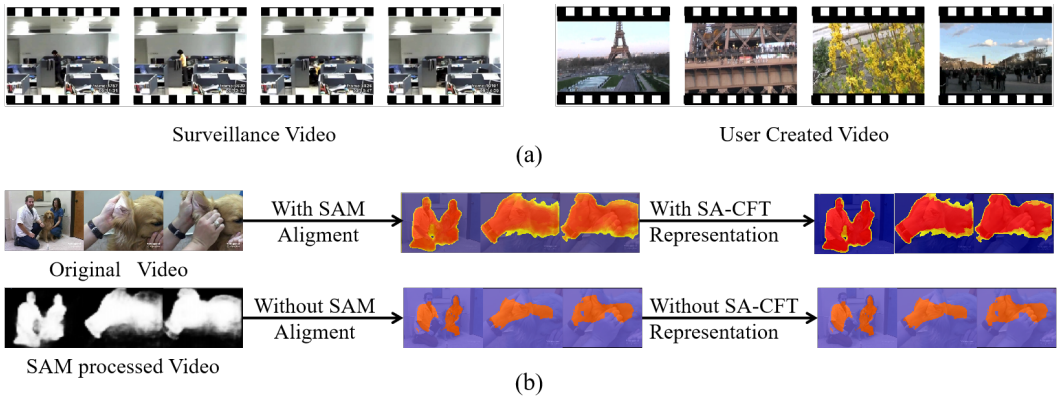

Fig. 1. (a) An example illustrating the difference between surveillance video and user-created video. While the surveillance video consists of a relatively static scene, the user-created video contains diversified content and more frequent shot changes. (b) The heatmap illustrating the enhancement of the discriminative ability in spatio-temporal domain by our contributions: 1) with/without the Spatial Attention Model (SAM) alignment, 2) with/without the Semantic representation and Attention alignment-based Contextual Feature Transformation (SA-CFT) representation; the more red the colour is, the stronger its discriminative ability, and vice versa.

addressed the issue of limited coding capability caused by the gate structure. Thus, the end-to-end LSTM model demonstrated superior performance in video summarization. Despite its success, the LSTM model has a principal weakness which is the poor generalization resulting from the over-fitting issue that generally occurs during training. The weakness leads to low precision in node classifications and then constrains the accuracy in video summarization.

Given the discriminative power of the Generative Adversarial Networks (GANs), Li proposed an unsupervised Cycle-Consistent Adversarial LSTM (Cycle-LSTM) by integrating a frame selector and a cycle-consistent learning-based evaluator [13]. The Cycle-LSTM method improved the performance of summarization over GAN-based LSTM [14]. Despite its strong capacity for representation learning [15], the performance of Cycle-LSTM drops with the unstable training of the cycle-adversarial component in video summarization. Alternatively, Graph Neural Networks (GNNs) had been successfully adopted for representation learning [16].

The Graph Attention Networks (GAT) [17] have proven to be effective in learning relationships between nodes by utilizing a self-attention mechanism. The model learns attention weights to aggregate feature vectors of nodes and their adjacent nodes, which enables efficient feature representation of each node. This method has demonstrated superior expressive and generalization abilities in graph data processing. Our previous work [18] proposed a GAT-based Bi-directional Long Short-Term Memory (Bi-LSTM) architecture to improve the representation of both node features and the graph structure simultaneously, and this approach was shown to effectively alleviate the issue of poor generalization in video summarization. However, the paper [18] revealed that the generalization capability of GAT can be undermined by noise in user-created video summarization.

Despite the strengths of the GAT, it can be sensitive to noise during graph node and edge construction in user-created video summarization. The GAT establishes an attention mechanism to assign variable weights to each neighbor for feature aggregation by learning the structure information among graph nodes [17]. However, recent studies have shown that deeper network layers can lead to an over-smoothing effect where node representations on stacked propagation become indistinguishable for samples of different classes [19]. Specifically, GAT's representation learning may map samples with non-identical semantic or visual features nearby, which can cause a decrease in video summarization performance. Moreover, the diverse data in user-created videos forms more complicated graph structures and node features, further limiting the efficiency of representation learning using GAT.

To address the issue of over-smoothing, we propose a Contextual Feature Transformation (CFT) mechanism that builds upon GIB [20] to enhance the temporal correlation among images. GIB is based on the Information Bottleneck (IB) [21] principle, which aims to learn the minimum sufficient representation of a given task by maximizing the mutual information between the representation and the output, while constraining the mutual information between the representation and the input data [20]. By doing so, GIB can strike a balance between the expressiveness and robustness of graph data representation. In our approach, we leverage the power of GIB to learn a more informative and discriminative representation of the input data, which is then further enhanced by our proposed CFT mechanism to capture the context and refine the temporal correlation among images.

In this study, we present a Semantic representation and Attention alignment-based Contextual Feature Transformation (SA-CFT) mechanism within the end-to-end Bi-LSTM framework. Our contributions are as follows:

1) To capture the diverse content in user-created videos, which is richer than surveillance videos, we introduce attention alignment to extend the representation from a single semantic feature space to a dual-feature space that includes both semantic and visual features.

2) To enhance the discriminative power and robustness of the dual-feature, we propose the SA-CFT mechanism, which refines the representation learning of the higher-layer dual-feature through stacked propagation of the GIB. As illustrated in Fig. 1(b), the SA-CFT mechanism emphasizes the discrim-

inative ability of the higher-layer dual-feature via temporal refinement.

3) Based on the observation that moving and large objects attract more attention, we introduce a Salient-Area-Size-based Spatial Attention Model (SAM) to refine the spatial features by leveraging the statistical information of the salient regions of moving objects. Fig. 1(b) shows that the SAM emphasizes the discriminative ability of the visual feature, enabling the proposed method to focus on the moving and large foreground.

4) To tackle the convergence problem during the training process, we employ the Deep Deterministic Policy Gradient (DDPG) algorithm [22]. Additionally, we use the binary cross-entropy loss function during supervised learning to obtain more reasonable training results.

In conclusion, our proposed method addresses issues of poor generalization and over-smoothing by refining the spatio-temporal representation of the dual-feature. We conduct comprehensive experiments on widely-used datasets, SumMe and TVSum, and demonstrate that our method outperforms SOTA methods [11] and [18] by 32.9%/8.5% and 6.8%/5.8% $F-score$ improvements on the SumMe/TVSum datasets, respectively.

II. Related Work

This section provides an overview of the classic visual attention model-based and deep neural network-based methods for video summarization. The deep neural network-based methods are further classified into two categories: single-feature and dual-feature methods. The single-feature methods operate on a single feature space, such as the semantic or visual feature space. On the other hand, the dual-feature methods extract both semantic and visual features, hence the name "dual-feature".

A. Classic visual attention model-based video summarization

Visual attention has been defined in the computer science literature since the mid-1990s [23], [9], and used as a criterion to measure the frame-wise importance in video summarization. Ji et al. [24] designed an iso-content principle-based saliency filter to select keyframes, which filters frames with lower importance scores than a predetermined threshold. Subsequently, Shih [4] generalized video summarization as a keyframe determination problem solved by ranking frame-wise attention levels. Since the attention is measured with hand-crafted features, such as block-wise temporal motion and facial area, Shih's method is constrained to specific video domains.

Even though the hand-crafted features correspond to limited representation ability, researchers have attempted to improve the performance of the classic methods by merging various visual cues, like moving objects, motion, foreground object categories in video summarization [5], [6]. Recently, Kannan presented a spatio-temporal saliency model to estimate the frame-wise importance [25], where the spatial-temporal saliency is developed based on low-level features such as color contrast, color distribution, and center prior. Kannan's method has shown superior performance to other classic visual attention model-based methods in video summarization.

Visual attention, as a low-level feature, was designed to represent videos with relatively static backgrounds or specific video domains in classic methods. However, extending visual attention cues to enhance the efficiency of these methods is limited in their ability to learn intricate patterns of large datasets. Classic methods fall short when dealing with the rapidly changing content in videos and fail to generate high-quality video summaries. Deep learning models, with their automatic multi-level representation power, have shown great potential in achieving high efficiency when addressing large datasets or general videos with complex content. Therefore, video summarization methods based on deep learning models have emerged as a promising approach to improve the performance of video summarization.

B. Spatio-temporal single-feature deep neural network-based video summarization

Supervised methods: Gong et al. [26] designed a novel probabilistic model, the Sequential Determinantal Point Process (SeqDPP), to learn the optimal subset of videos with the informative criteria in a supervised fashion. Besides, Zhang et al. [27] first adopted a LSTM to exploit the temporal dependencies, improving accuracy of video summarization by strengthening the keyframe selection criterion based on the work in [26].

Subsequently, Ji et al. [8] proposed an encoder-decoder framework with a novel attention-based LSTM mechanism for video summarization. In particular, the encoder-decoder framework cannot efficiently perform the keyframe selection using a fixed-length encoding vector for long video sequences. To address this, the attention-based LSTM mechanism was developed to assign the encoding vector to frames with more salient visual information. Ji emphasized the short-term contextual attention on the long-term attention model [8]. However, the supervised methods in video summarization depend heavily on the quality of the annotated dataset, which are highly labor-intensive to create.

Unsupervised methods: The discriminative ability of GAN has led to its adoption in video summarization. Mahasseni et al. [14] presented an approach with a variational auto-encoder and a novel adversarial LSTM network, where the auto-encoder selects keyframes, and the adversarial LSTM acts as a discriminator to differentiate between the original and reconstructed video frames. Li et al. [13] subsequently proposed an unsupervised Cycle-Consistent Adversarial LSTM (Cycle-LSTM) that combines a Bi-LSTM network for selecting keyframes and a cycle-consistent GAN structure to maximize the mutual information between the original and reconstructed videos. The Cycle-LSTM method improves the performance of summarization compared to Mahasseni's method [14]. However, the cycle-adversarial component suffers from unstable training results that lead to the model's failure in selecting keyframes.

Reinforcement learning in video representation simulated the essential mechanism of human annotation. Zhou et al. [11] presented an end-to-end LSTM-based reinforcement learning method for video summarization, which comprised a novel

feedback reward with the combination of diversity and representativeness. The method was proven to be successful by gradually enhancing the performance of video summarization during reinforcement learning. But, the idea that promotes video summarization by designing a more effective reward mechanism has its limitations [12].

The principal weakness of single-feature deep neural network-based methods is that the representation of a single features is often inadequate for discriminating between images with similar visual information but different semantic features. For instance, two images with similar visual information but completely different semantic features tend to be recognized as similar images in a class. They fail to identify the indistinguishable images as different, which causes missing detection of keyframes. To overcome this problem, researchers have proposed dual-feature methods, which involve embedding two types of features using deep learning models. By incorporating complementary features, these methods can enhance representation efficiency and improve the accuracy of video

C. Spatio-temporal dual-feature deep neural network-based unsupervised video summarization

Supervised methods: The work in [28] first introduced Convolutional Recurrent Neural Networks (CRNN) to exploit the spatio-temporal semantic feature-related dependencies and shallow features under an end-to-end architecture. More concretely, the combination of the semantic and shallow features made the representation used for video summarization to be more comprehensive, thereby producing superior performance against the single-feature methods, such as that in [11]. LMHA/LMHA-two [29] proposes a novel hierarchical attention approach for supervised video summarization that takes advantage of the inherent hierarchical structure of video sequences. It utilizes intra-block and inter-block attention mechanisms to learn both short-range and long-range temporal representations. Inspired by the success of transformer-based methods, HMT [30] introduces a hierarchical multimodal transformer architecture for video summarization that leverages both visual and textual information. The results from standard datasets demonstrate that both approaches outperform SOTA methods in terms of effectiveness and efficiency. However, due to the over-fitting phenomenon, which easily occurred when the annotated labels were insufficient, the performance of the supervised dual-feature methods decreased sharply.

Unsupervised methods: Based on Mahasseni's adversarial LSTM network [14], Jung et al. presented a variance loss regularization for discriminative feature learning under the Variational Auto Encoder-GAN (VAE-GAN) framework, where a novel Chunk and Stride Network (CSNet) was designed to combine local and global temporal video features [31]. The key idea was to enhance the performance of video summarization by integrating the semantic feature and the attention score measuring the temporal dynamic information. CSNet-based video summarization demonstrated superior performance compared to single-feature methods, showing that the dual-feature embedding approach improves the discriminative power of the representation for video summarization. Furthermore, this method was proved to be more efficient than the combination of shallow features and semantic features presented in [28].

Despite the enhanced performance, the work in [31] had two disadvantages. First, the attention score emphasized dynamic temporal differences rather than the visual features that truly represent the human visual perception mechanism. To overcome this limitation, our prior work [18] proposed a dual-feature method by integrating visual and semantic features. Specifically, the visual information captures the visual perception mechanism, wherein larger and moving objects attract more attention. Therefore, the combination of semantic and saliency features can further enhance the discrimination of algorithms for selecting keyframes and filtering redundant frames from user-created videos. Second, the adversarial components of the CSNet method [31] caused unstable training and significantly diminished performance. Therefore, in our previous work [18], we adopted an end-to-end LSTM model as the baseline network and leveraged GAT to enhance representation learning for video summarization.

D. Graph Neural Networks (GNNs)

In recent years, deep learning has been extended to graph-structured data [32], [33], resulting in the success of GNNs in various applications. GNNs learn node-level representations through message passing and neighboring aggregation, which helps in maintaining the topology information during the optimization process [34]. Graph Convolutional Neural Networks (GCNs), which generalizes the operation of Convolutional Neural Networks (CNNs) to graphs of arbitrary structures, have been proposed as an important branch of GNNs for representation learning in node classification [35], [36], [17]. Kipf proposed the original spectral-GCN for semi-supervised learning on graph data [35], which aimed to approximate and simplify the Chebyshev spectral CNN (ChebNet) [37] to make it more efficient. However, since the whole graph (including all nodes and edges) was processed and stored on the basis of Fourier transformation, the spectral-GCN still suffered from high computation and memory costs.

Spatial GCNs, such as GraphSage [36] and the Graph Attention Networks (GAT) [17], have been shown to be an efficient way to reduce the computational and memory costs for large-scale graphs. In spatial GCNs, convolution is directly performed on the graph, making it possible to perform training on a subset of nodes and edges within the graph. Hamilton proposed an inductive framework called GraphSage, which propagated information along the edges and aggregated features for a central node from its neighboring nodes [36]. The inductive framework predicted the structure's information by using the aggregated information for a node, facilitating the model's generalization capability. Moreover, GraphSage has been successfully applied to process complicated large-scale data, such as videos with millions of frames. However, GraphSage has a limitation in representing the dependency among nodes due to the inappropriate identical weight setting method that assigns the same weights to neighboring nodes even when they contribute differently to the current node.

GAT, on the other hand, assign variable weights to relevant neighboring nodes based on their contributions through the attention mechanism [17]. Compared to the identical weight assignment in GraphSage, GAT has shown impressive improvement in node classification. Considering GAT's strong generalization ability, our previous work [18] leveraged GAT to fuse the contribution of adjacent images to the current image in terms of visual features and then transform the visual features into higher-layer features. Furthermore, the GAT-based Bi-LSTM in [18] enhanced the representation of both graph structure and node features simultaneously, improving summary performance.

However, the representation learned by GAT is not robust and leaves features or structures vulnerable to disturbance [38], [39]. Recent work has claimed that the over-smoothing phenomenon that occurs during node representation on stacked propagation causes low efficiency in differentiating indistinguishable samples [19]. Particularly with the representation learning of GAT, samples with non-identical semantic or visual features tend to be mapped nearby, causing a sharp drop in the performance of video summarization for diversified data with more complicated graph structures and node features.

To make the graph more robust against disturbance, an information theory principle called the Graph Information Bottleneck (GIB) was presented in [20]. GIB learns the minimum sufficient representation of a given task by maximizing the mutual information between the representation and the output, and constraining the mutual information between the representation and the input data. Furthermore, GIB establishes an information theory model based on the characteristics, structure, and fusion of graph structure data. It focuses on the compression of node features and graph structures and improves prediction ability.

While dealing with large-scale and high-dimensional graph data, existing graph neural networks often face difficulties in interpretation and generalization. To address these issues, various methods have been proposed. For instance, Yu introduced the Graph Information Bottleneck (GIB) framework [40], which maps subgraphs in images to a low-dimensional space for efficient subgraph identification. However, this model suffers from unstable training and degraded results. To address these problems, Yu proposed the Variational Graph Information Bottleneck (VGIB) model [41], which combines the VAE model with the information bottleneck network to eliminate noise in the learning process and improve recognition accuracy.

In addition, Miao proposed the Graph Stochastic Attention (GSAT) method [42], which injects randomness to block task-irrelevant information and selects subgraph information relevant to the task by reducing randomness. However, random sampling may cause a loss of useful information. Sun proposed the Variational Information Bottleneck guided Graph Structure Learning framework (VIB-GSL) [43], which uses two neural networks to transform a graph into a probability distribution and embed it into a low-dimensional representation. The variational information bottleneck is used to limit the information flow of the embedded vector in the training process, thereby enhancing the ability to capture graph structure information.

In this work, we aim to enhance the discriminative ability of video summarization through the development of a Contextual Feature based Transformation mechanism using the GIB model. Our proposed mechanism generates a higher-layer dual-feature that is refined using stacked propagation of the GIB. Specifically, we focus on refining the representation learning of the temporal dual-feature, which includes semantic representation and attention alignment. To achieve this, we propose a semantic representation and attention alignment scheme for the GIB model under the end-to-end Bi-LSTM framework. Our approach enables effective summarization of videos with diverse content and complex graph structures.

III. Problem Formulation

Given a long video $\mathcal{X} = \{X_t\}_{t=1}^{T}$, where $X_t \in \mathbb{R}^{w \times h \times 3}$, and w, h, and 3 denote the width, height, and channel for each frame, respectively, $t \in \mathbb{N}^+$ is the index of the frame. T is the length of the video frames. $\mathbf{y} = \{y_t\}_{t=1}^{T}$, $y_t \in \{0, 1\}$ is the binary label representing whether or not the t_{th} frame is selected as a keyframe. The collection of the selected keyframes composes the video summary represented by $\mathcal{X}_{sub} = \{X_t \mid t < T,\ and\ y_t = 1\}$.

Generating Video Summarization: The binary possibility set $\mathbf{y} = \{y_t\}_{t=1}^{T}$ is computed on the Bernoulli distribution [36] of $\{\beta_t\}_{t=1}^{T}$ ($\beta_t \in [0, 1]$), as formulated in Eq. (1):

$$y_t = \text{Bernoulli}(\beta_t), \tag{1}$$

where $\beta_t = N(p_t)$ denotes the Soft Selected Probability (SSP) which is the probability of being selected as keyframes in video summarization, and $y_t = 1$ alternatively represents that the t_{th} frame with β_t is selected, and vice versa. $\mathrm{N}(\cdot) = \exp(\cdot)/\sum \exp(\cdot)$ is the normalization operator. p_t is the probability computed by the Bi-LSTM in Eq. (2):

$$p_t = \sigma(\mathrm{FC}(\mathbf{h}_t)), \tag{2}$$

where the hidden states $\{\mathbf{h}_t\}_{t=1}^{T}$ of the Bi-LSTM are inputted into a Fully Connected layer $\mathrm{FC}(\cdot)$ [11], followed by a sigmoid function $\sigma(\cdot)$ [44]. The input of the Bi-LSTM is the higher-layer feature M' calculated via a step of compression followed by a prediction within the SA-CFT.

IV. Proposed Method

As illustrated in Fig. 2, we propose a semantic representation and Spatial Attention Model's (SAM) alignment-based GIB model under the Bi-LSTM framework; the proposed model has an enhanced performance in discriminating between images by integrating semantic and visual features. First, a SAM is modeled by measuring the frame-wise spatial importance scores with the input of saliency features. Simultaneously, the semantic features are learned by a CNN [22]. Second, the concatenation of the semantic and spatial attention features is further transformed into higher-layer features by the SA-CFT established on the GIB. Lastly, the higher-layer features are processed by the Bi-LSTM network to generate the SSP. The reinforcement network, DDPG, is adopted to minimize the back-propagation loss.

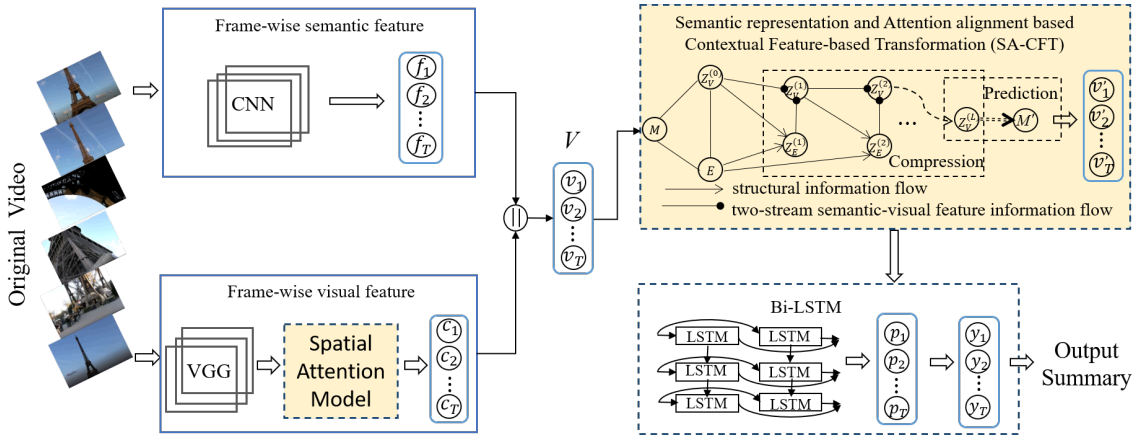

Fig. 2. Framework of the proposed method. Initially, we integrate the semantic features extracted from ImageNet with the visual features obtained through the Spatial Attention Model (SAM). Then, we feed the fused features into the SA-CFT network based on the Graph Information Bottleneck (GIB) architecture, where $Z_V^{(0)}$ is initialized as the inputted nodes' set $V = \{v_t\}_{t=1}^T$ with the fused feature v_t. The input features are then converted into higher level features. Lastly, we input the optimized features into the Bi-LSTM network to select key frames and generate a summary video. Our contributions consist of two key components: 1) the Graph Information Bottleneck (GIB) based SA-CFT network, and 2) the Spatial Attention Model (SAM) based attention alignment, combined with the Bi-LSTM network.

A. Semantic representation and Attention alignment-based Contextual Feature Transformation (SA-CFT)

To promote the discriminative ability of the dual-feature and improve its robustness, we develop the Semantic representation and Attention alignment-based Contextual Feature Transformation (SA-CFT) mechanism on the GIB model to generate a higher-layer dual-feature M. As shown in Fig. 2, the process of SA-CFT mechanism involves two steps, the compression followed by the prediction. The representation $Z_V^{(L)} = \{\mathbf{z}_{V,t}^{(L)}\}_{t=1}^T$ ($\mathbf{z}_{V,t}^{(L)} \in \mathbb{R}^d$ at the L_{th} iteration, and $d = d_1 + d_2$) is estimated during compression, and then the final output $M' = \{\mathbf{m}_t'\}_{t=1}^T$ ($\mathbf{m}_t' \in \mathbb{R}^d$) is computed via the prediction (see **Algorithm 1**).

We regard image frames as nodes set to generate an undirected attribute graph denoted as $G = (E, V)$, where $V = \{\mathbf{v}_t\}_{t=1}^T$ ($\mathbf{v}_t \in \mathbb{R}^d$) is the set of the dual-features. In Fig. 2, the semantic features $F = \{\mathbf{f}_t\}_{t=1}^T$ ($\mathbf{f}_t \in \mathbb{R}^{d_2}$, and d_2 is a different cardinality) and the visual features $C = \{\mathbf{c}_t\}_{t=1}^T$ ($\mathbf{c}_t \in \mathbb{R}^{d_1}$, $d_1 = w \times h$) are concatenated to form the input V, modelled as Eq. (3):

$$\mathbf{v}_t = [\mathbf{f}_t \parallel \mathbf{c}_t], \tag{3}$$

where $\parallel$ is the concatenation operator. While the semantic feature $\mathbf{f}_t$ is learned via a pre-trained GoogleNet on ImageNet [22], the visual feature $\mathbf{c}_t$ is generated using the SAM model. Let $E = \{e_{i,j}^{(l)}\}$ be the graph's adjacency matrix ($l \in \mathbb{N}$, and $l \leq L+1$ is the iteration time). In our approach, we initialize the graph's adjacency matrix E using Eq. (4), where $e_{i,j}^{(0)} = 1$ if nodes i and j belong to the same clip and $e_{i,j}^{(0)} = 0$ otherwise. More specific, there is an edge between nodes i and j if and only if they belong to the same clip. Conversely, if i and j belong to different clips, they are not connected, and $e_{i,j}^{(0)}$ equals 0.

$$e_{i,j}^{(0)} = \begin{cases} 1, & i,j \in \textit{the same clip} \\ 0, & \text{else} \end{cases}. \tag{4}$$

The KTS algorithm [45] is a widely used algorithm for segmenting time series data into multiple clips. The main goal of this algorithm is to maximize the similarity between time series data within each clip while minimizing the similarity between different clips, thereby improving analysis and processing. In the context of video processing, KTS [45] is employed to segment videos into multiple clips, where each clip comprises a number of video frames, denoted by the variable k. Within each clip, the output for each frame, denoted by $\mathbf{m}_i$, with $i \in [1, k]$, is calculated through the prediction in Eq. (5). The output M is obtained by combining the outputs $\{\mathbf{m}_i\}_{i=1}^k$ ($\mathbf{m}_i \in \mathbb{R}^d$) of all clips.

$$\mathbf{m}_i = \sum\nolimits_{\mathrm{j}\in[1,\mathrm{k}],\mathrm{j}\neq\mathrm{i}} N_j(e_{ij}^{(l)}) \cdot (U\mathbf{z}_{V,j}^{(l-1)}) \text{ s.t. } l = L+1, \tag{5}$$

where the weight $e_{ij}^{(l)}$ of the edge between the i_{th} and j_{th} nodes is computed on a shared attention mechanism δ, $U \in \mathbb{R}^{d\times d}$ is a matrix with learnable weights, and $N_j(\cdot)$ is the normalization with the index j. Based on the GAT [17], the model of attention weight $e_{ij}^{(l)}$ can be formulated as Eq. (6):

$$e_{ij}^{(l)} = \delta(U\mathbf{z}_{V,i}^{(l-1)}, U\mathbf{z}_{V,j}^{(l-1)}), \tag{6}$$

where the attention weight $e_{ij}^{(l)}$ presents the interaction impact of the j_{th} node's feature $\mathbf{z}_{V,j}^{(l-1)}$ on the i_{th} node's feature $\mathbf{z}_{V,i}^{(l-1)}$ at the l_{th} iteration.

The equation of the attention weight is further expressed in Eq. (7) by introducing the LeakyReLU nonlinearity and a weight vector $\mathbf{a} \in \mathbb{R}^{2d}$,

$$e_{ij}^{(l)} = \mathrm{LeakyReLU}(\mathbf{a}^T[U\mathbf{z}_{V,i}^{(l-1)} \parallel U\mathbf{z}_{V,j}^{(l-1)}]), \tag{7}$$

where $\mathbf{a}^T$ is the transposition of the vector $\mathbf{a}$.

The basic framework of GIB-based SA-CFT is shown in **Algorithm 1**. The stacked propagation of the GIB is used to refine the representation learning of the temporal dual-feature (semantic representation with attention alignment) for video summarization. The GIB learns the minimum sufficient

Algorithm 1: The Framework of GIB-based SA-CFT

Input: The dataset $G = (E, V)$;
k: The number of neighbors to be sampled;
Output: $Z_V'^{(L)}$,
$M = \mathcal{W}_{out} \odot Z_V^{(L)} = \{W_i^{(L+1)} \mathbf{z}_{V,i}^{(L)}\}_{i=1}^{k}$,
$M' = \mathcal{W}'_{out} \odot Z_V'^{(L)} = \{W_i'^{(L+1)} \mathbf{z}'^{(L)}_{V,i}\}_{i=1}^{k}$

Initialize: $Z_V^{(0)} \leftarrow V$; for all $\mathbf{v}_t \in V$,
Weights: $U \in \mathbb{R}^{d\times d}$, $W_i^{(l)}, W_i'^{(l)} \in \mathbb{R}^{d\times d}$,
$\mathcal{W}_{out} = \{W_i^{(L+1)}\}_{i=1}^{k}$, $\mathcal{W}'_{out} = \{W_i'^{(L+1)}\}_{i=1}^{k}$,
$W_i^{(l)} = \sum_{j\in[1,k],j\neq i} N_j(e_{ij}^{(l)}) \cdot U$ for $l \in [1, L+1]$
$W_i'^{(l)} = \sum_{j\in[1,k],j\neq i} N_j(e_{ij}'^{(l)}) \cdot U$ for $l = L+1$
a. The upper bound of $\mathrm{I}(G; Z_V^{(L)})$
for *iteration* $l = [1, \cdots, L]$, **do**
  i) calculate $\mathrm{VIB}^{(l)}$:
    $\mathbf{z}_{V,i}^{(l)} \leftarrow W_i^{(l)} \mathbf{z}_{V,i}^{(l-1)}$; for all $i \in [1, k]$
    $\mu_v^{(l)} \leftarrow \{\mathbf{z}_{V,i}^{(l)}\}_{i=1}^{k/2}$
    $\{\sigma_v^2\}^{(l)} \leftarrow \mathrm{softplus}(\{\mathbf{z}_{V,i}^{(l)}\}_{i=k/2}^{k})$
    variable $z_v^{(l)} \sim Gaussian(\mu_v^{(l)}, \{\sigma_v^2\}^{(l)})$
    $\mathrm{VIB}^{(l)} = \sum_{z_v^{(l)} \in Z_V^{(l)}} \big[\log\Phi(z_v^{(l)}; \mu_v^{(l)}, \{\sigma_v^2\}^{(l)})$
    $-\log(\sum_{n=1}^{100} \lambda_n^{(l)} \cdot \Phi(z_v^{(l)}; \mu_n^{(l)}, \{\sigma_n^2\}^{(l)}))\big]$,
  ii) calculate $\mathrm{EIB}^{(l)}$:
    $\phi_v^{(l)} = \mathrm{softmax}([\mathbf{z}_{V,i}^{(l-1)} \parallel \mathbf{z}_{V,j}^{(l-1)}]\mathbf{a}^T)$
    $\mathrm{EIB}^{(l)} = \sum_{z_v^{(l)} \in Z_V^{(l)}} \mathrm{KL}(\mathrm{Cat}(\phi_v^{(l)}) || \mathrm{Cat}(z_v^{(l)}))$
  iii) $Z_V^{(l)} = \min(\mathrm{VIB}^{(l)})$ and $Z_E^{(l)} = \min(\mathrm{EIB}^{(l)})$
  iv) update $e_{ij}^{(l)}$:
    $e_{ij}^{(l)} = \mathrm{LeakyReLU}(\mathbf{a}^T[U\mathbf{z}_{V,i}^{(l-1)} \parallel U\mathbf{z}_{V,j}^{(l-1)}])$
end
b. The lower bound of $\mathrm{I}(M; Z_V^{(L)})$
$\mathrm{I}(M; Z_V^{(L)}) \geq \mathrm{MIB} = -\sum_{v\in V} \mathrm{L_s}(\mathcal{W}_{out} \odot Z_V^{(L)}, Z_V^{(L)})$
c. By $\min(-\mathrm{MIB} + \gamma_1 \mathrm{VIB}^{(L)} + \gamma_2 \mathrm{EIB}^{(L)})$, we update to obtain the optimal $Z_V'^{(L)} = \{\mathbf{z}'^{(L)}_{V,i}\}_{i=1}^{k}$, and update $e_{ij}'^{(L+1)} = \mathrm{LeakyReLU}(\mathbf{a}^T[U\mathbf{z}'^{(L)}_{V,i} \parallel U\mathbf{z}'^{(L)}_{V,j}])$.
d. generate the prediction
$M' = \mathcal{W}'_{out} \odot Z_V'^{(L)} = \{W_i'^{(L+1)} \mathbf{z}'^{(L)}_{V,i}\}_{i=1}^{k}$

representation of a keyframe selection task by maximizing the mutual information $\mathrm{I}(M; Z_V^{(L)})$ between the representation $Z_V^{(L)}$ and the output M, and constraining the mutual information $\mathrm{I}(G; Z_V^{(L)})$ between the representation $Z_V^{(L)}$ and the input data G, as modelled in Eq. (8):

$$\underset{P(Z_V^{(L)}|G)\in\Omega}{\mathrm{argmin}} \mathrm{GIB}_\gamma(G, M; Z_V^{(L)}) = \quad [-\mathrm{I}(M; Z_V^{(L)}) + \gamma \mathrm{I}(G; Z_V^{(L)})], \tag{8}$$

where $P(Z_V^{(L)} \mid G)$ denotes the Probability Distribution Functions (PDFs) of the random variable $z_v^{(L)}$ within the set $Z_V^{(L)}$ given the condition $G = (E, V)$, and $\gamma \in \{\gamma_1, \gamma_2\}$ are the hyperparameters. However, graph-structured data is known to suffer from the problem of being non-i.i.d (non-independent and identically distributed). Because of the Markov correlation between video frames, it can be assumed that the graph data has local dependence on the adjacent nodes' set. Conversely, the current node is assumed to be independent of the nodes beyond the adjacent nodes' set. Therefore, we restrict the search space Ω to the adjacent nodes' set. To simplify the optimization process for GIB, we employ variational bounds [46] to develop and optimize the terms $\mathrm{I}(M; Z_V^{(L)})$ and $\mathrm{I}(G; Z_V^{(L)})$.

The upper bound of $\mathrm{I}(G; Z_V^{(L)})$: minimizes the mutual information between representation $Z_V^{(L)}$ and the input information G, which are used to eliminate the disturbance of noise data in the inputted features, calculated in Eq. (9):

$$\begin{aligned}\mathrm{I}(G; Z_V^{(L)}) &\leq \mathrm{I}(G; \{Z_V^{(l)}\}_{l\in[1,l_v]} \cup \{Z_E^{(l)}\}_{l\in[l_v+1,L]}) \\ &\leq \sum_{l\in[1,l_v]} \mathrm{VIB}^{(l)} + \sum_{l\in[l_v+1,L]} \mathrm{EIB}^{(l)},\end{aligned} \tag{9}$$

where $\mathrm{VIB}^{(l)}$ and $\mathrm{EIB}^{(l)}$ represent the mutual information between the updated representation and the original input in terms of feature and correlation structure respectively, and $l_v \in [1, L]$. At the l_{th} iteration, $Z_V^{(l)}$ is the updated representation of the dual-feature, and $Z_E^{(l)}$ denotes the representation of the structure information between the current node and its $(k-1)$ adjacent nodes.

We calculate $\mathrm{VIB}^{(l)}$ in Eq. (10):

$$\begin{aligned}\mathrm{VIB}^{(l)} &= E\left[\log\frac{P(Z_V^{(l)} \mid Z_V^{(l-1)}, Z_E^{(l)})}{Q(Z_V^{(l)})}\right] \\ &= \sum_{z_v^{(l)}\in Z_V^{(l)}} \big[\log\Phi(z_v^{(l)}; \mu_v^{(l)}, \{\sigma_v^2\}^{(l)}) \\ &\quad -\log(\sum_{n=1}^{100} \lambda_n^{(l)} \cdot \Phi(z_v^{(l)}; \mu_n^{(l)}, \{\sigma_n^2\}^{(l)}))\big],\end{aligned} \tag{10}$$

where $P(Z_V^{(l)} \mid Z_V^{(l-1)}, Z_E^{(l)})$ indicates the PDFs of the variable $z_v^{(l)} \sim Gaussian(\mu_v^{(l)}, \{\sigma_v^2\}^{(l)})$ for the set $Z_V^{(l)}$ given the conditions $Z_V^{(l-1)}$ and $Z_E^{(l)}$ at the l_{th} iteration. $E[\cdot]$ is the expectation value of the random variables. $Q(Z_V^{(l)})$ denotes the PDFs of the Gaussian Mixture Model for set $Z_V^{(l)}$ after reparameterization (the corresponding prior distribution is written as $z_v^{(l)} \sim \sum_{n=1}^{100} \lambda_n^{(l)} \cdot Gaussian(\mu_n^{(l)}, \{\sigma_n^2\}^{(l)})$ [47]. $Gaussian(:,:)$ is the probability density function of the Gaussian distribution, denoted by $\Phi(:;:,:)$. We set $\lambda_n^{(l)}$, $\mu_n^{(l)}$, and $\{\sigma_n^2\}^{(l)}$ as learnable parameters.

As for $\mathrm{EIB}^{(l)}$ in Eq. (11), we set $P(Z_E^{(l)} \mid E, Z_V^{(l-1)})$ as the PDFs of the variable $z_e^{(l)} \sim \mathrm{Cat}(\phi_v^{(l)})$ in the set $Z_E^{(l)}$ given the conditions E and $Z_V^{(l-1)}$, and set $Q(Z_E^{(l)})$ as the PDFs of the prior distribution $\mathrm{Cat}(z_v^{(l)})$ for set $Z_E^{(L)}$. Specifically, the KL divergence [48] $\mathrm{KL}(\mathrm{Cat}(\phi_v^{(l)})||\mathrm{Cat}(z_v^{(l)}))$ is leveraged to measure the similarity between the conditional distribution $\mathrm{Cat}(\phi_v^{(l)})$ and the prior distribution $\mathrm{Cat}(z_v^{(l)})$.

$$\begin{aligned}\mathrm{EIB}^{(l)} &= E\left[\log\frac{P(Z_E^{(l)} \mid E, Z_V^{(l-1)})}{Q(Z_E^{(l)})}\right] \\ &= \sum_{z_v^{(l)}\in Z_V^{(l)}} \mathrm{KL}(\mathrm{Cat}(\phi_v^{(l)})||\mathrm{Cat}(z_v^{(l)})),\end{aligned} \tag{11}$$

where $\phi_v^{(l)} = \mathrm{softmax}([\mathbf{z}_{V,i}^{(l-1)} \parallel \mathbf{z}_{V,j}^{(l-1)}]\mathbf{a}^T)$ indicates the importance of the features of the j_{th} node to the i_{th} node at the l_{th} iteration. $\mathrm{Cat}(\phi_v^{(l)})$ is the Gumbel-Softmax [49] function of $\phi_v^{(l)}$ with temperature τ at the l_{th} iteration. $\mathrm{Cat}(z_v^{(l)}) = |Z_V^{(l)}|^{-1}$ is the uniform distribution for the nodes' set $Z_V^{(l)}$ with the structure $Z_E^{(l)}$.

In summary, we can update $Z_V^{(l)}$ and $Z_E^{(l)}$ by $Z_V^{(l)} = \min(\mathrm{VIB}^{(l)})$ and $Z_E^{(l)} = \min(\mathrm{EIB}^{(l)})$, respectively.

The lower bound of $\mathrm{I}(M; Z_V^{(L)})$: maximizes the mutual information between the output M and the representation $Z_V^{(l)}$, calculated in Eq. (12):

$$\begin{aligned}\mathrm{I}(M; Z_V^{(L)}) \geq 1 + E\left[\log \frac{\prod_{z_v^{(l)} \in Z_V^{(l)}} Q(M \mid Z_V^{(L)})}{Q(M)}\right] \\ + E_{P(M)P(Z_V^{(L)})}\left[\frac{\prod_{z_v^{(l)} \in Z_V^{(l)}} Q(M \mid Z_V^{(L)})}{Q(M)}\right].\end{aligned} \tag{12}$$

The last term of Eq. (12) equals 1 empirically. We simply set $Q(M \mid Z_V^{(L)}) = \mathrm{Cat}(\mathcal{W}_{out} \odot Z_V^{(L)})$ and $Q(M) = P(M)$. Thus, the right side of Eq. (12), denoted by MIB, is deduced as Eq. (13),

$$\mathrm{MIB} = -\sum_{z_v^{(l)} \in Z_V^{(l)}} \mathrm{L_s}(\mathcal{W}_{out} \odot Z_V^{(L)}, Z_V^{(L)}), \tag{13}$$

$\mathrm{L_s}(:,:) = \mathrm{L}(\mathrm{sigmoid}(:),:)$ is the combination of a binary cross-entropy loss $L(:,:)$ (formulated in Eq. (18)) and a sigmoid function $\mathrm{sigmoid}(:)$ [50]. Furthermore, $\odot$ is element-wise product with the first order of the tensor $\mathcal{W}_{out}$ and the corresponding vector $\mathbf{z}_{V,i}^{(l-1)}$, modelled in **Algorithm 1**. Last, we update to obtain the optimal $Z_V^{\prime(L)}$ via $\min(-\mathrm{MIB} + \gamma_1 \mathrm{VIB}^{(L)} + \gamma_2 \mathrm{EIB}^{(L)})$, where $\gamma_1 = 0.0001$ and $\gamma_2 = 0.001$ are the hyperparameters of $\mathrm{VIB}^{(L)}$ and $\mathrm{EIB}^{(L)}$, respectively.

B. Salient-Area-Size-based SAM

Inspired by the idea that human beings frequently pay more attention to the sizable moving objects in a video, we propose a saliency feature-based model, SAM, to formulate spatial attention to the statistics of the saliency region of moving objects. Let $S = \{\mathbf{s}_t \mid \mathbf{s}_t \in [0, 255]\}_{t=1}^{T}$ denote the set of saliency map. In Fig. 3, the core salient region detection is proposed to refine the saliency map $\mathbf{s}_t \in \mathbb{R}^{d_1}$ to generate the frame-wise spatial attention feature $\mathbf{c}_t \in \mathbb{R}^{d_1}$ ($t \in [1, T]$, $d_1 = w \times h$) in Eq. (14):

$$\mathbf{c}_t = \begin{cases} \mathbf{s}_t, & r_\partial > r_{\hat{\partial}} \\ 0, & \text{else} \end{cases}, \tag{14}$$

where $\mathbf{s}_t$ is the vector representing the pixel-wise salient value. Specifically, a feature vector is extracted by a VGG [11], on which a saliency detector [23] is adopted to detect the salient regions of moving objects.

As for the pixel-wise saliency map $\mathbf{s}_t$, we rank the area of salient regions and then set the salient regions whose area r_∂ is less than the threshold $r_{\hat{\partial}}$ as non-salient regions.

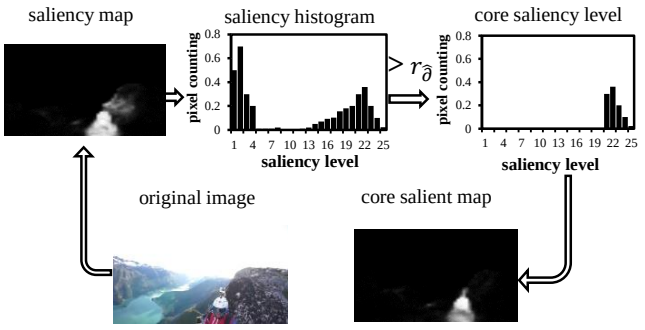

Fig. 3. Spatial attention model based on salient regions. This model is utilized to identify the key salient regions within an image by identifying the regions with significance levels greater than the threshold value.

To calculate the optimal value of $r_{\hat{\partial}}$ for determining the core salient regions, we conduct experiments by ranging $r_{\hat{\partial}}$ from 20 to 25 and randomly sampling 150 videos for testing. In Fig. 4, when $r_{\hat{\partial}} = 21$, SAM arrives at the best performance in $F{-}score$. A saliency map histogram is rendered from the saliency map $\mathbf{s}_t$, whose horizontal coordinate is the saliency level r_∂, while the vertical coordinate is the frequency $\mathrm{bins}(\cdot)$. The range of horizontal coordinates is 26 saliency levels presented as $\{r_0, r_1, \cdots, r_\partial, \cdots, r_{25}\}$, $\partial \in [0, 25]$, where the class interval of each level is 10 (expect the special case $r_{25} = 5$). In the saliency histogram, the frequency of vertical coordinates is computed by the discrete function $\mathrm{bins}(\cdot)$, written as Eq. (15):

$$\mathrm{bins}(r_\partial) = \frac{o_{r_\partial}}{w \times h}, \tag{15}$$

where o_{r_∂} is the number of pixels at the salient regions with the saliency level r_∂.

C. Training-Unsupervised fashion (DDPG)

We adopt the DDPG [22] to train the SA-CFT-based Bi-LSTM. The DDPG is a united algorithm of the actor-critic deterministic policy gradient algorithm [22], which contains two kinds of models: the actor and the critic. The actor takes action according to the environment state, and the critic evaluates the actor's action and provides action-value. In this work, the actor is the SA-CFT-based Bi-LSTM. The proposed method serves as the actor to learn a policy μ_{θ^μ} by maximizing the expected actor-value, written as Eq. (16), and then a liner regression network is developed to play as a critic.

$$\max_{\theta^\mu} J(\mu) = E_{s \sim p^\beta}\left[Q^{\mu_{\theta^\mu}}\left(s, \mu_{\theta^\mu}(s)\right)\right], \tag{16}$$

where θ^μ is the parameter of the policy, p^β is the state visitation distribution, s is the hidden state in the network, and $Q^{\mu_{\theta^\mu}}(\cdot)$ is the actor-value function, modelled by the critic. Following [22], we adopt the deterministic policy gradient algorithm to calculate the derivative of $J(\mu)$. The policy gradient is formulated as Eq. (17):

$$\bigtriangledown_{\theta^\mu} J \approx \sum_i \bigtriangledown_a Q\big(s, a \mid \theta^Q\big)|_{s=s_i, a=\mu(s_i)} \bigtriangledown_{\theta^\mu} \mu\big(s \mid \theta^\mu\big)|_{s_i}, \tag{17}$$

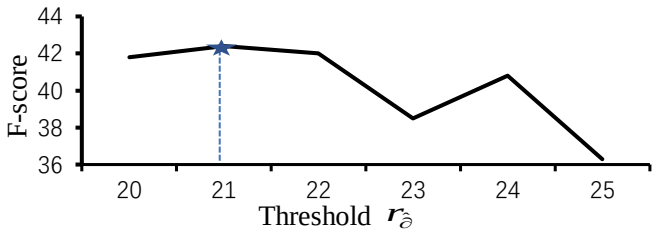

Fig. 4. Illustration depicting the $F-score$ function of with respect to the threshold value. Its primary objective is to determine the optimal threshold for identifying the core saliency region in an image.

where the parameters in the critic are updated by minimizing the Smooth-L1 Loss $J(\mu)$ between the predicted and the expected action-value.

D. Training-Supervised fashion

To compare with the supervised methods fairly, we extend our model into a supervised fashion with the binary cross-entropy loss $\mathrm{L}(:,:)$ in Eq. (18),

$$\mathrm{L}(\beta_t, g_t) = -\frac{1}{k}\sum_{t=1}^{k}\Big[g_t \log(\beta_t) + (1-g_t)\log(1-\beta_t)\Big], \quad (18)$$

where β_t represents the predicted SSP modelled in Eq. (1), g_t is the SSP of the annotated ground-truth, and k is the length of the video frames within a clip.

V. Experiments

A. Implementation Details

Datasets: We evaluate our model on two public datasets: SumMe [51] and TVSum [52], where SumMe consists of 25 user-created videos covering a wide range of topics (such as sports, vacations, and holidays) and TVSum contains 50 edited videos in 10 categories selected from YouTube. The video durations range from 1 to 10 minutes. More than 15 users give frame-level importance scores to create manually annotated summaries. Most of the videos from TVSum are edited versions rather than the original ones. Two other datasets, YouTube [53] and the Open Video Project (OVP), are adopted to augment the training process [8], [11]. The YouTube dataset contains 39 videos selected from the YouTube channel, excluding cartoon videos, and the OVP consists of 50 documentary videos. Generally, the duration of the algorithm-generated summary should be less than 15% of that of the original video.

Evaluation metrics: We adopt the commonly used evaluation metric, the $F-score$ [27] which assesses performance by measuring the similarity between the algorithm-generated summary A and the gold-standard labels B annotated by the users. First, the precise P and the recall R are calculated by Eq. (19), and then the harmonic mean $F-score$ is computed in Eq. (20):

$$\begin{aligned} P &= \frac{\text{overlapped duration of } A \text{ and } B}{\text{duration of } A}, \\ R &= \frac{\text{overlapped duration of } A \text{ and } B}{\text{duration of } B}, \end{aligned} \quad (19)$$

$$F-score = 2 \times \frac{P \times R}{P + R} \times 100\%. \quad (20)$$

Experimental settings: We conducted experiments on three settings, namely Canonical (C), Augmented (A), and Transfer (T), as listed in Table I. In the Canonical setting, we utilized a single dataset and allocated 80% of it for training and 20% for evaluation. In the Augmented setting, we employed four distinct datasets, with 80% of the first dataset and three others for training, and 20% of the first dataset for testing. In the Transfer setting, we used three datasets for training and one dataset for evaluation. Our experimental methodology followed the parameter configuration outlined in [8], [11], which included down-sampling the video to 2 frames per second and dividing it into multiple smaller segments. The hidden layer size of the Bi-LSTM was set to 256. During the intensive learning training, we set the episode to 5 and the learning rate to $1e-4$. Furthermore, we set the maximum number of iterations during network training to 60 and recorded changes in the loss every 5 epochs. We performed validation experiments with five rounds and reported the averaged $F-score$.

TABLE I
Experimental settings for SumMe and TVSum under the Canonical (C), Augmentation (A), and Transfer (T) Settings, respectively.

Setting	Training set	Test set
Canonical	80%TVSum	20%TVSum
Augmented	80%TVSum+YouTube+OVP+SumMe	20%TVSum
Transfer	TVSum+YouTube+OVP	SumMe

B. Quantitative Results

Comparison with multiple network structures: To evaluate the performance of the proposed summarization method comprehensively, we present the impact of multiple network structures, such as Gate Recurrent Unit (GRU) [54], Recurrent Neural Network (RNN) [55], CNN [56], and Bi-LSTM [44], on the experimental results. In Fig. 2, we illustrate the case that the Bi-LSTM [44] network is combined with the proposed SA-CFT and SAM models. Moreover, Bi-LSTM can be replaced by the GRU [54], RNN [55], or CNN [56].

TABLE II
Comparison among multiple network structures in an unsupervised fashion under the canonical setting ($F-score\%$).

Method	SumMe	TVSum
GRU [54]	54.1	61.4
RNN [55]	54.3	61.8
CNN [56]	54.8	62.5
Bi-LSTM	**55.0**	**62.5**

As shown in Table II, we can see that different network structures have a particular impact on the experimental results. The GRU performs the worst. Compared to the Bi-LSTM, it is with 1.66% and 1.79% loss in $F-score$ on SumMe and TVSum, respectively. Moreover, the CNN network shows potential performance with only 0.36% loss in $F-score$ against the Bi-LSTM on SumMe. However, the Bi-LSTM network achieves the best performance among the above networks. Therefore, we adopt the Bi-LSTM as our baseline network.

TABLE III
ABLATION STUDY. COMPARISON OF THE SUMMARY PERFORMANCE OF VARIOUS MODELS IN THE PROPOSED FRAMEWORK USING UNSUPERVISED LEARNING UNDER THE CANONICAL SETTING ($F-score\%$).

mode	DR-DSN	Bi-LSTM+DDPG	GAT	SA-CFT	SAM	$F-score\%$	
						SumMe	TVSum
1	√					41.4	57.6
2		√				45.9	57.8
3		√			√	50.9	58.5
4		√	√			51.1	58.7
5		√		√		52.6	62.4
6		√	√		√	51.5	59.1
7		√		√	√	**55.0**	**62.5**

Ablation study: In Table III, we utilize mode 1 to mode 7 to denote the baseline method DR-DSN [11] (mode 1), our previous work [18] (mode 4 and 6), and the proposed models (mode 2, 3, 5, and 7). The proposed method comprises the GIB-based SA-CFT mechanism for transforming the dual-feature into a higher-layer feature and the Salient-Area-Size-based SAM for spatial feature refinement.

The ablation results of the proposed method on unsupervised training justify that both the SA-CFT mechanism and SAM model can enhance performance efficiently. Mainly, for the SumMe and TVSum datasets, the SAM (mode 3) and the SA-CFT (mode 5) perform at 10.89% and 1.21% gain (5.0% and 0.7% absolute gain), as well as 14.6% and 7.96% gain (6.7% and 4.6% absolute gain) compared to the baseline network of the proposed method (mode 2). Moreover, the combination of the SAM and the SA-CFT (mode 7) raises the 19.83% and 8.13% $F-score$ against mode 2 (9.1% and 4.7% absolute gain).

Furthermore, the proposed method outperforms the DR-DSN (mode 1) significantly. The SAM (mode 3) and the SA-CFT (mode 5) achieve improvements of 22.95% and 1.56% gain (9.5% and 0.9% absolute gain) and 27.05% and 8.33% gain (11.2% and 4.8% absolute gain). Simultaneously, the combination (mode 7) arrives at 32.85% and 8.51% gain (13.6% and 4.9% absolute gain) against mode 1. The superior performance of the SAM (mode 3), the SA-CFT (mode 5), and the combination (mode 7) demonstrates that the proposed method effectively alleviates the weak generalization in DR-DSN by refining the dual-feature.

In addition, compared to our previous method [18] (mode 6), the proposed method (mode 7) reaches 6.80% and 5.75% gain (3.5% and 3.4% absolute gain). The improvements verify that the SA-CFT is more robust than the GAT by generating a higher-layer spatio-temporal feature. Meanwhile, we see that the SA-CFT and the SAM are compatible in enhancing the performance.

In our proposed SA-CFT, we integrate the Contextual Feature Transformation (CFT) to refine the representation learning of higher-layer dual-features by leveraging semantic representation and attention alignment. To assess the efficiency of the CFT, we further utilize it to refine the temporal correlation among single-feature, semantic feature, or visual feature processed via the SAM, referred to as S-CFT and A-CFT respectively.

Additionally, in the S-CFT, A-CFT, and SA-CFT models, we can substitute the Gaussian Mixture Model (GMM-100) with a Gaussian model for the distribution of $Z_V^{(l)}$. An ablation study was performed to compare the performance of the Gaussian Model and Gaussian Mixture Model. The SA-CFT model utilizing GMM-100 exhibits an improvement in $F-score$ of 3.38% and 0.15% (an absolute gain of 1.8% and 0.9%) compared to the Gaussian distribution. On the other hand, when using the Gaussian distribution, A-CFT performs slightly better than the S-CFT on the TVSum dataset, indicating the potential to enhance video summarization through visual features.

TABLE IV
ABLATION STUDY. COMPARISON OF THE SUMMARY PERFORMANCE OF S-CFT, A-CFT, AND SA-CFT MODELS USING UNSUPERVISED LEARNING UNDER THE CANONICAL SETTING ($F-score\%$).

CFT	models	SumMe	TVSum
S-CFT	Gaussian	52.7	61.2
	GMM-100	53.4	61.7
A-CFT	Gaussian	52.5	61.4
	GMM-100	-	-
SA-CFT	Gaussian	53.2	61.6
	GMM-100	55.0	62.5

Comparison with unsupervised approaches: We select 12 approaches as our baselines and then compare these methods with our model on SumMe and TVSum. The baselines can be roughly categorized into three classes: 1) the spatial structure-based methods include K-medoids [57], Vsumm [53], Web image [58], Dictionary selection [57], Online space coding [59], and Co-archetypal [52]; 2) the temporal structure-based methods contain GANdpp [14], DR-DSN [11], Cycle-SUM [13], and CSNet [31]; 3) the spatio-temporal structure-based method is the GAT adjusted Bi-LSTM [18] named GAT-LSTM , and the deep semantic and attention network (DSAVS) [60].

The experiments are conducted in an unsupervised fashion in the canonical setting [27]. As shown in Table V, the proposed method yields 32.85% and 8.51% gain (13.6% and 4.9% absolute gain) against DR-DSN on SumMe and TVSum. Compared to Cycle-SUM [13], the proposed method produces 31.26% and 8.51% gain (13.1% and 4.9% absolute gain), respectively. Besides, the proposed method outperforms CSNet [31], with 7.21% and 6.29% gain (3.7% and 3.7% absolute gain). The experimental results confirm that the proposed models overcome the unstable training of the adversarial components in Cycle-SUM and CSNet. More specifically, another limitation of the CSNet is that it presents an attention score by measuring the dynamic temporal difference, which

TABLE V
COMPARISON WITH OTHER UNSUPERVISED VIDEO SUMMARIZATION METHODS UNDER THE CANONICAL SETTING ($F-score\%$).

Method	SumMe	TVSum
K-medoids [57]	33.4	28.8
Vsumm [53]	33.7	-
Web image [58]	-	36.0
Dictionary selection [57]	37.8	42.0
Online space coding [59]	-	46.0
Co-archetypal [52]	-	50.0
GANdpp [14]	39.1	51.7
DR-DSN [11]	41.4	57.6
Cycle-SUM [13]	41.9	57.6
CSNet [31]	51.3	58.8
GAT-LSTM [18]	51.5	59.1
DSAVS [60]	47.0	59.4
Proposed	**55.0**	**62.5**

is proven to be inefficient in representing the visual feature of the video. In contrast, our SAM model presents genuine visual features based on the specified perception mechanism: humans prefer to focus on sizable and moving objects.

Our previous work in [18] adopted the end-to-end LSTM model as a baseline network and leveraged GAT to enhance the representation learning for video summarization. However, the proposed method also outperforms the GAT-LSTM [18] with 6.80% and 5.75% gain. The results have proven that GIB can generate a more robust high-layer dual-feature (semantic representation with attention alignment) than GAT, with better discriminative ability for video summarization. Compared to DSAVS [60], our method is with 17.02% and 5.22% gain (8.0% and 3.1% absolute gain) on SumMe and TVSum, respectively.

Moreover, it is worth noticing that the $F-score$ of the SumMe dataset is relatively lower for most comparison algorithms [57], [14], [11], [31] than those of the TVSum dataset, which proves that the baselines could not perform well with raw videos. However, the gain of a 32.9% $F-score$ (13.6% absolute gain) against DR-DSN on the SumMe dataset demonstrates that the proposed method alleviates this phenomenon.

Comparison with supervised approaches: The proposed method is compared to other 17 typical baselines in a supervised fashion on SumMe and TVSum. The experiments are conducted in the Canonical setting [27] and the results are listed in Table VI.

Compared to methods without temporal structure, such as Interestingness [51], ours improves the $F-score$ by 42.13% (16.6% absolute gain) on the SumMe dataset. Compared to methods based on temporal structure, such as Bi-LSTM [27], Dpp-LSTM [27], and GANsup [14], the proposed method also shows superior performance with a margin.

Furthermore, the proposed method increases the $F-score$ by 33.02% (13.9% absolute gain) on SumMe and 7.92% (4.6% absolute gain) on TVSum against the DR-DSNsup [11]. The proposed method attains 25% and 7.92% gain (11.2% and 4.6% absolute gain) against the Cycle-SUMsup [13]. Due to the unstable training, CSNetsup in a supervised version performs worse than the method in an unsupervised version. The proposed method outperforms the CSNetsup [31], with 15.23% and 7.18% gain (7.4% and 4.2% absolute gain). The proposed method has been proven to be more stable in training than Cycle-SUMsup and CSNetsup.

TABLE VI
COMPARISON WITH OTHER SUPERVISED VIDEO SUMMARIZATION METHODS UNDER THE CANONICAL SETTING ($F-score\%$).

Method	SumMe	TVSum
Interestingness [51]	39.4	-
Submodularity [61]	39.7	-
Summary transfer [62]	40.9	-
Bi-LSTM [27]	37.6	54.2
Dpp-LSTM [27]	38.6	54.7
GANsup [14]	41.7	56.3
DR-DSNsup [11]	42.1	58.1
Cycle-SUMsup [13]	44.8	58.1
CSNetsup [31]	48.6	58.5
CRSum [28]	47.3	58.0
GAT-LSTMsup [18]	51.7	59.6
HMT [30]	44.1	60.1
DSAVSsup [60]	48.9	59.8
DSNet [63]	50.2	62.1
LMHA [29]	51.1	61.0
LMHA-two [29]	51.4	61.5
RR-STG [64]	53.4	**63.0**
Proposed-sup	**56.0**	62.7

In contrast to the GAT-LSTMsup [18], ours improves the $F-score$ by 8.32% and 5.20% (4.3% and 3.1% absolute gain). Our proposed method, the SA-CFT, outperforms several transform-based methods including HMT [30], LMHA [29], and LMHA-two [29], on both the SumMe and TVSum datasets. Additionally, it achieves better results compared to DSAVSsup [60] and DSNet [63] with 14.52% and 11.55% $F-score$ gain (7.1% and 5.8% absolute gain) and 4.85% and 0.97% gain (2.9% and 0.6% absolute gain) on TVSum, respectively. However, compared to RR-STG [64], our method performs slightly worse on TVSum, with an absolute loss of 0.3%. Despite this, our method outperforms almost all of the supervised methods. These results suggest that the SA-CFT model, built on the GIB framework, can effectively address the over-smoothing issue that occurs during node representation in GAT. Specifically, the SA-CFT improves video summarization performance by enhancing the differentiation of indistinguishable samples.

TABLE VII
COMPARISON WITH OTHER VIDEO SUMMARIZATION METHODS UNDER THE CANONICAL (C), AUGMENTATION (A), AND TRANSFER (T) SETTINGS, RESPECTIVELY ($F-score\%$).

Method	SumMe			TVSum		
	C	A	T	C	A	T
Bi-LSTM [27]	37.6	41.6	40.7	54.2	57.9	56.9
Dpp-LSTM [27]	38.6	42.9	41.8	54.7	59.6	58.7
GANdpp [14]	39.1	43.4	-	51.7	59.5	-
GANsup [14]	41.7	43.6	-	56.3	61.2	-
DR-DSN [11]	41.4	42.8	42.4	57.6	58.4	57.8
DR-DSNsup [11]	42.1	43.9	42.6	58.1	59.8	58.9
CSNet [31]	51.3	52.1	45.1	58.8	59	59.2
CSNetsup [31]	48.6	48.7	44.1	58.5	57.1	57.4
GAT-LSTM [18]	51.5	53.6	42.8	59.1	59.0	57.1
GAT-LSTMsup [18]	51.7	52.2	43.8	59.6	59.5	57.9
HMT [30]	44.1	44.8	-	60.1	60.3	-
LMHA [29]	51.1	52.1	45.4	61.0	61.5	55.1
LMHA-two [29]	51.4	52.8	45.6	61.5	62.8	56.7
Proposed	55.0	54.0	42.9	62.5	60.4	57.4
Proposed-sup	56.0	54.8	44.0	62.7	60.3	58.0

Comparison with Canonical, Augmentation, and Transfer Settings: In the Canonical and Augmentation settings, the proposed method outperforms previous outstanding work by a large margin shown in Table VII. Since CSNet utilizes global and local information, the proposed method produces a slightly worse $F-score$ by 5.13% and 3.14% (2.2% and 1.8% absolute gain) on SumMe and TVSum with unsupervised learning in the Transfer setting. However, in a supervised fashion, the proposed method produces nearly the same performance as CSNetsup on SumMe and slightly better performance (0.6% absolute gain) than CSNetsup on TVSum.

Our proposed method outperforms the supervised method, HMT [30], LMHA [29], and LMHA-two [29], in both Canonical and Augmentation settings on the SumMe dataset. These results demonstrate that our method can alleviate issues caused by poor generalization or over-smoothing by refining the input data through temporal refinement of the dual-features via the SA-CFT. Compared to LMHA-two [29], our method achieves better results in the Canonical and Transfer settings but performs slightly worse in the Augmentation setting. This suggests that our model exhibits slightly poorer learning ability than a multi-scale hierarchical attention model in dealing with edited videos under different data characteristic conditions. In the Transfer setting, our method performs slightly worse than in the Canonical and Augmentation settings. Therefore, we plan to address this issue in future work.

Time complexity comparison: The proposed framework was implemented on a machine equipped with an Intel® Core^{TM} i9-10900K CPU, NVIDIA GeForce RTX 3090, and 32GB of RAM, running a 64-bit Ubuntu 18.04.05 LTS Operating System. The time complexity is evaluated with five rounds of experiments conducted on Seq1—Seq5 (each Seq includes ten videos randomly selected from the TVSum or SumMe datasets). In Table VIII, we present the time cost comparison between the proposed method and two baselines, DR-DSN [11] and GAT-adjusted Bi-LSTM [18]. The time cost per video is computed and tabulated in the last row of Table VIII.

TABLE VIII
COMPARISON OF MODELS' TIME COMPLEXITY USING UNSUPERVISED LEARNING (TIME: SECOND).

Sequence	DR-DSN [11]	GAT-LSTM [18]	Proposed
Seq1	1.2337	2.7705	2.5584
Seq2	2.4160	5.2917	5.3485
Seq3	2.3674	4.9910	5.1761
Seq4	1.7709	3.8535	3.7807
Seq5	1.8126	3.8478	4.0096
Average	0.1920	0.4151	0.4175

The proposed method comprises two components, a GIB followed by a Bi-LSTM. The GIB is a specific GNNs in time complexity, written as $O(|V| \cdot d \cdot d + |E| \cdot d)$. d is the cardinality of the dual-feature input feature, and $|V|$ and $|E|$ are the numbers of nodes and edges for the video frames, respectively. Moreover, the time complexity of the Bi-LSTM equals the sum of complexities of two LSTM layers, $2{\cdot}O(n{\cdot}d)$, where n is the size of the hidden layer.

Since the GAT and GIB correspond to similar time complexity, the time complexity of both the proposed method and the GAT-adjusted Bi-LSTM [18] is approximate $O(|V| \cdot d^2 + |E| \cdot d) + 2 \cdot O(n \cdot d)$. By contrast, the time complexity of the DR-DSN is low with only a Bi-LSTM model $2 \cdot O(n \cdot d)$. In Table VIII, the GAT-adjusted Bi-LSTM [18] and the proposed method cost $0.2231s$ and $0.2255s$ more time per video than DR-DSN, respectively. However, the proposed method achieves a significantly higher $F-score$ than reference methods in video summarization, with impressive $F-score$ improvements as high as 32.85% and 8.51% gain (13.6% and 4.9% absolute gain) against DR-DSN [11].

C. Quality Results

Selected Keyframes: In Fig. 5, we visualize the keyframes selected via the proposed method and user annotations with red and gray bars, respectively, for four different videos (a)-(d) on the TVSum dataset. It can be seen that most high-score frames are selected by the proposed method. The visualization results indicate that the proposed method can accurately predict the keyframes approaching human selections via the combination of the SA-CFT mechanism and the SAM model.

Spatial Importance: In Fig. 6, the spatial importance scores calculated via the SAM and the corresponding user-labelled scores are labelled by red and gray bars, respectively. The spatial importance score α_t is computed as the sum of saliency histograms of the core salient regions in Eq. (21):

$$\alpha_t = \sum_{r_\partial > r_{\hat{\partial}}}^{r_{25}} \text{bins}(r_\partial). \tag{21}$$

The trend of the blue curve presents the distribution of the frame-wise salient regions' area, which follows the distribution of the user-labelled scores. It demonstrates that the SAM can extract the visual features obeying human perception. Furthermore, we illustrate the spatial attention feature maps for four frames with the highest Spatial Importance below the bar chart. The visualization of the spatial importance scores has shown that the semantic information is compatible with the visual information.

D. Failure case

In Fig. 7, we illustrate a failure case. The inputted video "Saving Dolphins" on SumMe can be divided into two clips: 1) dolphins are washed onto the beach, labelled as Event 1; 2) people on the beach help dolphins return to the sea, labelled as Event 2. The moving objects (like the dolphins) are captured from far to near in Event 1. Since the camera shot is taken from a long distance, the dolphins are too small to be recognized as an attention object. The problem is that the summary generated by the proposed method only contains Event 2. The proposed model is likely to fail when the moving object is far away from the shooting place, as in the sample "Saving Dolphins". The lack of long-distance scene detection via SAM causes failure in this kind of case. We plan to learn a model that can integrate long-distance scene detection in the future in this field.

VI. CONCLUSION

This paper introduces a new approach to address the challenges of poor generalization and over-smoothing in video

(a)video_5 (b)video_18

(c)video_41 (d)video_42

Fig. 5. Visualization of comparison between the proposed method and the user-annotated ground-truth for four videos (a)-(d) from the TVSum. The gray bars are the user-annotated frame-wise scores, and the red bars denote the keyframes' SSP (predicted by the proposed method).

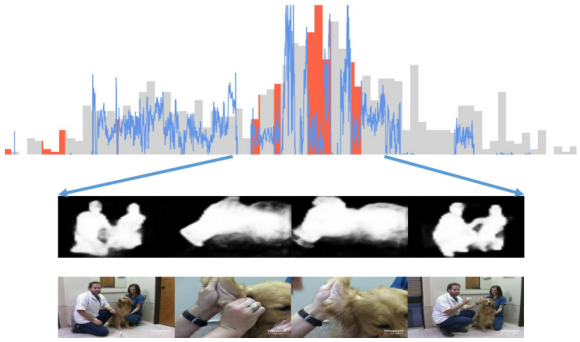

Fig. 6. Visualization of the spatial importance scores. The gray bars denote the user-annotated ground-truth, and the red bars indicate the spatial importance scores of the keyframes predicted by the Spatial Attention Model (SAM). An extensive blue curve is the frame-wise salient regions' area plotted as a comparison to the user-labelled score distribution. We illustrate the spatial attention feature maps for four frames with the highest Spatial Importance shown below the bar chart.

summarization. The proposed method incorporates attention alignment and spatial attention models to create a dual-feature space that includes both visual and semantic features. The SA-CFT mechanism uses the GIB stacked propagation to refine the representation learning of the dual-feature, enhancing its discriminative power and robustness. To overcome the convergence problem during training, we employed the DDPG algorithm and binary cross-entropy loss function in supervised learning. Our experiments on SumMe and TVSum datasets showed that our method outperformed the SOTA approaches. We also found that the SAM model had limited accuracy in detecting telephoto targets due to the domain adaption issue under the Transfer setting in a supervised fashion. To address this issue, we plan to integrate a telephoto target detection module into the SAM model.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62002130, 62201222), and in part by the Fundamental Research Funds for the Central Universities (CCNU22QN014, CCNU22XJ034 and CCNU22JC007), in part by the National Key Research and Development Program of China (2022YFD1700204), and in part by Fonds Wetenschappelijk Onderzoek (FWO), Vlaanderen, under Project G094122N.

References

[1] J. Han, K. Li, L. Shao, X. Hu, S. He, L. Guo, J. Han, and T. Liu, "Video abstraction based on fmri-driven visual attention model," *Information Sciences (IS)*, vol. 281, pp. 781–796, 2014.

[2] G. Wu, J. Han, Y. Guo, L. Liu, G. Ding, Q. Ni, and L. Shao, "Unsupervised deep video hashing via balanced code for large-scale video retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1993–2007, 2018.

[3] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "Rgb-t salient object detection via fusing multi-level cnn features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2019.

[4] H.-C. Shih, "A novel attention-based key-frame determination method," *IEEE Trans. Broadcast.*, vol. 59, no. 3, pp. 556–562, 2013.

[5] M. M. Salehin and M. Paul, "Summarizing surveillance video by saliency transition and moving object information," in *International Conference on Digital Image Computing: Techniques and Applications*. IEEE, 2015, pp. 1–8.

[6] H. Jacob, F. L. Pádua, A. Lacerda, and A. C. Pereira, "A video summarization approach based on the emulation of bottom-up mechanisms of visual attention," *Journal of Intelligent Information Systems*, vol. 49, no. 2, pp. 193–211, 2017.

[7] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *Computer Science*, 2015.

[8] 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, 2019.

[9] 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, 2020.

[10] Z. Ji, F. Jiao, Y. Pang, and L. Shao, "Deep attentive and semantic preserving video summarization," *Neurocomputing*, vol. 405, pp. 200–207, 2020.

[11] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *AAAI Conf. Artif. Intell.*, 2018, pp. 7582–7589.

[12] N. Gonuguntla, B. Mandal, N. Puhan, *et al.*, "Enhanced deep video summarization network," in *British Machine Vision Conference*, 2019, pp. 1–9.

[13] L. Yuan, F. E. H. Tay, P. Li, and J. Feng, "Unsupervised video summarization with cycle-consistent adversarial lstm networks," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2711–2722, 2019.

[14] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 202–211.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 35, no. 8, pp. 1798–1828, 2013.

[16] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, 2021.

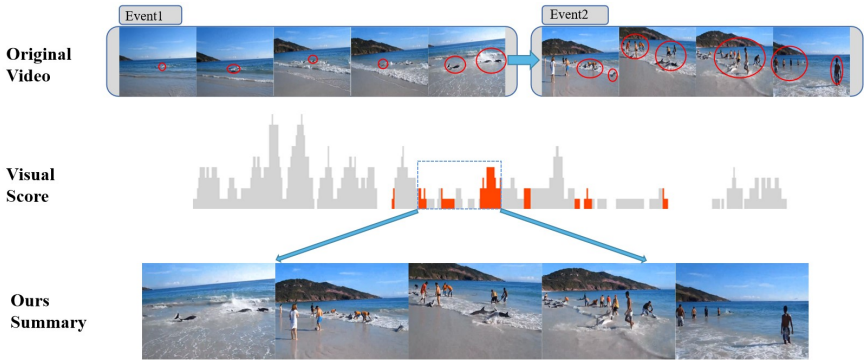

Fig. 7. Visualization of a failure case, "Saving Dolphins". The first row, "Original Video", indicates the two parts of the "Saving Dolphins" video, where the red circle annotates moving objects. The second row, "Visual Score", shows the comparison between the proposed method (red bars) and the user-annotated ground-truth (gray bars). The gray bars are the user-annotated frame-wise scores and the red bars denote the SSP (predicted by the proposed method) of the keyframes. The third row, "Ours Summary", is predicted by the proposed method, whose content only contains the original video's Event 2.

[17] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. of the Int. Conf. Learn. Represent.*, 2018.

[18] R. Zhong, R. Wang, Y. Zou, Z. Hong, and M. Hu, "Graph attention networks adjusted bi-lstm for video summarization," *IEEE Signal Process. Lett.*, vol. 28, pp. 663–667, 2021.

[19] M. Liu, H. Gao, and S. Ji, "Towards deeper graph neural networks," in *Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 338–348.

[20] T. Wu, H. Ren, P. Li, and J. Leskovec, "Graph information bottleneck," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20 437–20 448.

[21] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[22] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *Computer Science*, 2015.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. of the Int. Conf. Learn. Represent.*, 2014.

[24] Q.-G. Ji, Z.-D. Fang, Z.-H. Xie, and Z.-M. Lu, "Video abstraction based on the visual attention model and online clustering," *Signal Process. Image Commun.*, vol. 28, no. 3, pp. 241–253, 2013.

[25] R. Kannan, S. Swaminathan, G. Ghinea, F. Andres, and K. S. M. Anbananthen, "Movie video summarization-generating personalized summaries using spatiotemporal salient region detection," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 10, no. 3, pp. 1–26, 2019.

[26] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 2069–2077, 2014.

[27] 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.

[28] Y. Yuan, H. Li, and Q. Wang, "Spatiotemporal modeling for video summarization using convolutional recurrent neural network," *IEEE Access*, vol. 7, pp. 64 676–64 685, 2019.

[29] W. Zhu, J. Lu, Y. Han, and J. Zhou, "Learning multiscale hierarchical attention for video summarization," *Pattern Recognition*, vol. 122, p. 108312, 2022.

[30] M. Sanabria, F. Precioso, and T. Menguy, "Hierarchical multimodal attention for deep video summarization," in *International Conference on Pattern Recognition*, 2021, pp. 7977–7984.

[31] Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon, "Discriminative feature learning for unsupervised video summarization," in *AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 8537–8544.

[32] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proc. IEEE International Joint Conference on Neural Networks*, vol. 2, 2005, pp. 729–734.

[33] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[34] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Int. Conf. Mach. Learn.*, 2017, pp. 1263–1272.

[35] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. of the Int. Conf. Learn. Represent.*, 2017.

[36] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1025–1035.

[37] D. Michal, B. Xavier, and V. Pierre, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3844–3852.

[38] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2847–2856.

[39] H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song, "Adversarial attack on graph structured data," in *Int. Conf. Mach. Learn.*, 2018, pp. 1115–1124.

[40] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He, "Graph information bottleneck for subgraph recognition," in *Proc. of the Int. Conf. Learn. Represent.*, 2021.

[41] J. Yu, J. Cao, and R. He, "Improving subgraph recognition with variational graph information bottleneck," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 19 396–19 405.

[42] S. Miao, M. Liu, and P. Li, "Interpretable and generalizable graph learning via stochastic attention mechanism," in *Int. Conf. Mach. Learn.*, 2022, pp. 15 524–15 543.

[43] Q. Sun, J. Li, H. Peng, J. Wu, X. Fu, C. Ji, and S. Y. Philip, "Graph structure learning with variational information bottleneck," in *AAAI Conf. Artif. Intell.*, vol. 36, no. 4, 2022, pp. 4165–4174.

[44] W. Zhong, H. Xiong, Z. Yang, and T. Zhang, "Bi-directional long short-term memory architecture for person re-identification with modified triplet embedding," in *Proc. Int. Conf. Image Process.*, 2017, pp. 1562–1566.

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

[46] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *Int. Conf. Mach. Learn.*, 2019, pp. 5171–5180.

[47] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," in *Proc. of the Int. Conf. Learn. Represent.*, 2017.

[48] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[49] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. of the Int. Conf. Learn. Represent.*, 2016.

[50] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Curran Associates Inc., 2018, p. 8792–8802.

[51] 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.

[52] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5179–5187.

[53] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[54] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proc. of the Conf.on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734.

[55] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," in *Proc. of the Int. Conf. Learn. Represent.*, 2014.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1–9, 2012.

[57] 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 Recog.*, 2012, pp. 1600–1607.

[58] 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 Recog.*, 2013, pp. 2698–2705.

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

[60] S.-H. Zhong, J. Lin, J. Lu, A. Fares, and T. Ren, "Deep semantic and attentive network for unsupervised video summarization," *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 2, pp. 1–21, 2022.

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

[62] 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 Recog.*, 2016, pp. 1059–1067.

[63] 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.

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

Rui Zhong (Member, IEEE) is an associate professor at the School of Computer Science, Central China Normal University (CCNU) since 2018. Prior to this, she was a post-doctoral researcher at the Electronics and Informatics (ETRO) department of the Vrije Universiteit Brussel (VUB), Belgium. She obtained her Bachelor's degree in Electronic Information School from Wuhan University, Wuhan, China in 2008, and her Doctorate degree in Computer Science from Wuhan University, China in 2014. Her research interests focus on video and 3D graphics coding and processing, as well as multimedia transmission.

Rui Wang is a current graduate student at the School of Computer Science, Central China Normal University (CCNU), China. Her primary research areas include video summarization, light field compression, and 3D image synthesis.

Wenjin Yao is a current graduate student at the School of Computer Science, Central China Normal University (CCNU), China. His primary research areas include object tracking, machine learning, and image classification.

Min Hu is a graduate student at National Engineering Research Center for Multimedia Software, Wuhan University, China. Her primary research areas include microexpression recognition, video summarization, and intelligent surveillance system.

Shi Dong is an associate professor at the Faculty of Artificial Intelligence in Education, Central China Normal University (CCNU) since 2017. He obtained his Bachelor's degree in the Department of Electronics and Information Engineering from Huazhong University of Science and Technology, Wuhan, China in 2008, and his Doctorate degree in Computer Science from Wuhan University, China in 2014. His research interests include 3D video and audio signal processing and natural language processing.

Adrian Munteanu (M'07) received the M.Sc. degree in electronics and telecommunications from Politehnica University of Bucharest, Romania, in 1994, the M.Sc. degree in biomedical engineering from University of Patras, Greece, in 1996, and the Doctorate degree (magna cum laude) in applied sciences from Vrije Universiteit Brussel (VUB), Belgium, in 2003. From 2004 to 2010, he was a Post-Doctoral Fellow with the Fund for Scientific Research-Flanders, Belgium and, since 2007, he has been a Professor at VUB. He is a Professor with the Department of Electronics and Informatics, VUB. He has authored over 400 journal and conference publications, book chapters, and contributions to standards and holds seven patents in image and video coding. His research interests include image, video and 3D graphics coding, distributed visual processing, 3D graphics, error-resilient coding, multimedia transmission over networks, and statistical modeling. He was a recipient of the 2004 BARCO-FWO Prize for his Ph.D. work, and of several prizes and scientific awards in international journals and conferences. He served as an Associate Editor for IEEE Transactions on Multimedia and currently serves as Associate Editor for IEEE Transactions on Image Processing.