Expert Systems With Applications 249 (2024) 123568

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Attention-guided multi-granularity fusion model for video summarization

Yunzuo Zhang *, Yameng Liu, Cunyu Wu

School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China

ARTICLE INFO

Keywords:
Video summarization
Multi-granularity
Content-aware enhancement
Scale-adaptive fusion
Self-attention
Temporal convolution

ABSTRACT

Video summarization has attracted extensive attention benefiting from its valuable capability to facilitate video browsing. While achieving notable improvement, existing methods still fail to sufficiently and effectively model contextual information within videos, hindering the summarization performance owing to a deficiency in powerful contextual representations. To address this limitation, we present a novel Attention-Guided Multi-Granularity Fusion Model (AMFM), which allows for optimizing the modeling process from the context capturing and fusion perspective. AMFM comprises three dominant components including a content-aware enhancement (CAE) module, a multi-granularity encoder (MGE), and a scale-adaptive fusion (SAF) module. More specifically, CAE dynamically enhances pre-trained visual features by learning the potential visual relationship across frame-level and video-level embeddings. Subsequently, coarse-grained and fine-grained contextual information is simultaneously modeled in the same representation space by MGE with the combination of self-attention and temporal convolution scheme. Furthermore, the multi-granularity representations with a significant difference in the semantic scale are adaptively fused by SAF. Our method can precisely pinpoint key segments by effectively modeling and processing rich temporal representations. Extensive comparisons with state-of-the-art methods on standard datasets demonstrate the effectiveness of the proposed method, and the ablation studies further verify the positive impact of each module in our model.

1. Introduction

Recently, the proliferation of mobile devices has led to an exponential increase in the number of videos (Hussain et al., 2021; Lin, Zhao, Su, Wang, & Yang, 2018), as evidenced by the staggering amount of videos uploaded to YouTube daily (James, 0000). This creates an urgent demand for intelligent video analysis, and video summarization has become a hot research topic aimed at reducing this overload. At its core, video summarization involves understanding the content of videos and generating a concise yet comprehensive synopsis by removing massive redundant content (Xiao, Zhao, Zhang, Guan, & Cai, 2020). To date, it has been studied in many specific scenarios (Bettadapura, Pantofaru, & Essa, 2016; Li, Pan, Wang, Xing, & Han, 2022; Merler et al., 2019; Xu et al., 2021; Zhang, Zhu and Roy-Chowdhury, 2016).

Video summarization can be broadly categorized into static methods and dynamic methods (Huang & Wang, 2020; Yuan, Mei, Cui, & Zhu, 2019). Static methods aim to select a set of key frames, while dynamic methods pick several key shots composed of consecutive frames to represent the entire video content. This paper concentrates on key shot-based video summarization, as dynamic summarization is more helpful for users to understand the video storyline.

Existing methods have made unprecedented progress in summarizing videos (Jiang & Mu, 2022; Liu et al., 2022; Park, Lee, Kim, & Sohn, 2020; Xie et al., 2022; Zhu, Lu, Li, & Zhou, 2021). The majority of traditional methods (De Avila, Lopes, da Luz Jr, & de Albuquerque Araújo, 2011; Gygli, Grabner, Riemenschneider, & Gool, 2014; Zhang, Tao, & Wang, 2017) concentrate on the selection of meaningful segments with heuristic representations based on hand-crafted features. Nevertheless, these features are limited in their ability to provide rich semantic information, and the temporal cues within videos are rarely exploited, which is insufficient for comprehensively understanding the source video. Recently, deep learning-based methods (Jung, Cho, Kim, Woo, & Kweon, 2019; Yuan, Tay, Li, & Feng, 2020; Zhu, Lu, Han, & Zhou, 2022) have gained increasing interest. Usually, lots of methods adopt Recurrent Neural Networks (RNNs) to enrich visual features with contextual information. For instance, Zhang, Chao, Sha, and Grauman (2016b) fed frame-level representations into Long Short-Term Memory (LSTM) for long-range temporal aggregation. Although these variants of RNNs are capable of effectively modeling the video sequence, they encounter gradient vanishing. Additionally, the difficulty in implementing parallel computing (Liang, Lv, Li, Zhang and Zhang, 2022) is also an aspect that cannot be ignored.

To address these issues, the fully convolutional sequence network (Rochan, Ye, & Wang, 2018) is proposed. However, it cannot effectively learn the pairwise relationship across frames. To improve the

* Corresponding author.
E-mail addresses: zhangyunzuo888@sina.com (Y. Zhang), liuym4647@sina.com (Y. Liu), wucunyu1410@sina.com (C. Wu).

https://doi.org/10.1016/j.eswa.2024.123568
Received 20 January 2024; Received in revised form 20 February 2024; Accepted 24 February 2024
Available online 27 February 2024

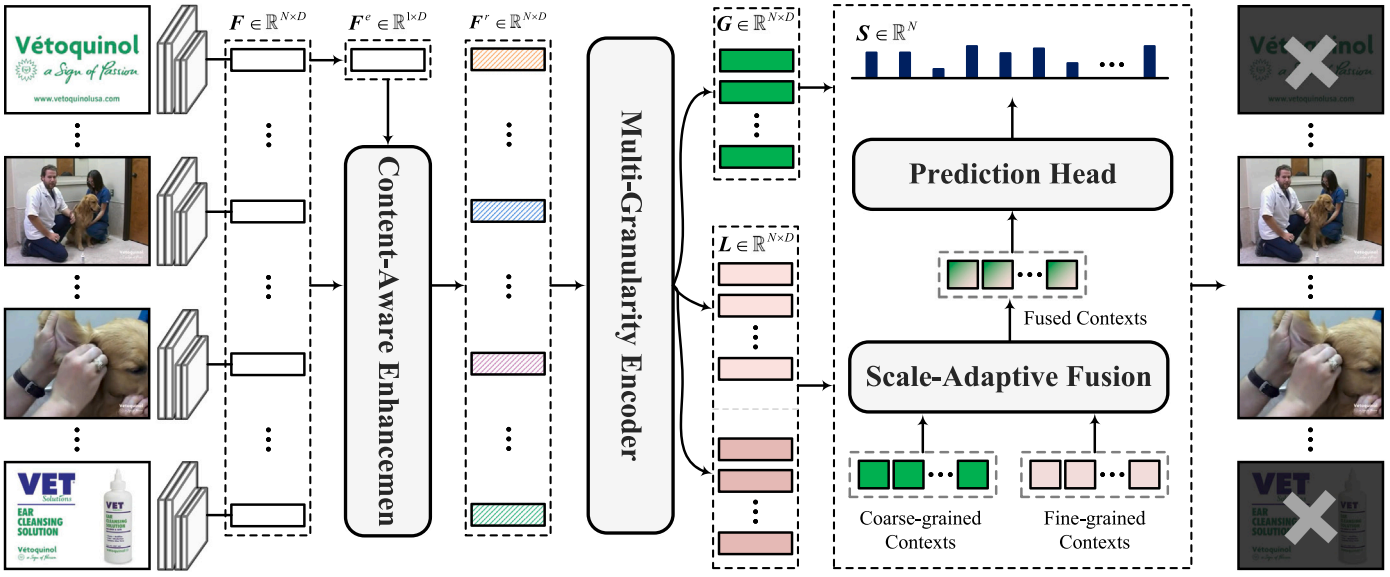

Fig. 1. Overview of AMFM. The visual features modeled from the input video are sequentially fed into the content-aware enhancement module, the multi-granularity encoder, and the scale-adaptive fusion module. Based on the fused contexts, the prediction head outputs importance scores, which are used for generating video summarization.

capability of video understanding, some attention-based hierarchical methods (Zhao, Gong and Li, 2022; Zhu et al., 2022) are proposed to model local contextual information before learning global contextual information. They usually segment the entire video sequence into subsequences, which serve as local modeling units and the basis for learning long-range contextual information (Zhao, Li, & Lu, 2017). However, such a rough division either introduces too much noise or discards too many valuable characteristics. Although shot boundary detection (Potapov, Douze, Harchaoui, & Schmid, 2014) is also exploited in existing efforts, they are still susceptible to inaccurate subsequence partitioning and cascading negative effects on global contextual feature learning due to the limited detection performance (Zhao, Li, & Lu, 2018). Actually, global and local contextual information can provide coarse-grained and fine-grained semantic information about the input video, respectively, allowing the deep learning-based model for comprehensive video understanding. Because of defective learning schemes in existing methods, contextual information within videos is still difficult to sufficiently and effectively model, hindering the summarization performance.

Given the aforementioned problem, this paper proposes a novel Attention-Guided Multi-Granularity Fusion Model (AMFM), which allows for optimizing the modeling process from the context capturing and fusion perspective. As shown in Fig. 1, AMFM consists of three dominant components including a content-aware enhancement (CAE) module, a multi-granularity encoder (MGE), and a scale-adaptive fusion (SAF) module. Firstly, based on the attention scheme, CAE targets to achieve feature enhancement by estimating the potential visual relationship across frame-level and video-level embeddings, learning powerful visual representations to facilitate the effectiveness of modeling contextual information. As an individual module, it can be easily embedded into our model to be trained in an end-to-end manner. Secondly, MGE simultaneously aggregates global and local contextual information with the combination of self-attention and temporal convolution scheme. Different from previous hierarchical structure-based methods, our method models entire video sequences and subsequences separately in a parallel manner, which avoids the potential negative impact of cascading learning. In addition, MGE can capture rich and finer temporal cues using multiple convolution operators for accurate video understanding.

After global and local contextual information is aggregated, SAF adaptively fuses these features by learning fusion attention. This is motivated by the fact that global and local contextual information is obtained from frames of significantly different ranges, hence intuitive fusion strategies cannot ensure powerful contextualized representations owing to the lack of deep interaction. Finally, the fused features are fed into a prediction head for frame-level importance score prediction and summary generation. In practice, the training procedure of the proposed method can be easily parallelized, making it computationally efficient. Extensive experiments on standard datasets are conducted, and empirical studies demonstrate the effectiveness of the proposed method.

In a nutshell, the main contributions of this paper can be summarized as follows:

- We develop the CAE module grounded on the attention scheme to learn powerful visual representations to facilitate the effectiveness of modeling contextual information within videos.
- We build the MGE module, which simultaneously models global and local contextual information and allows for robust learning of finer temporal cues for accurate video understanding.
- We consider the significant semantic scale difference across global and local contextual information, devising the SAF module to form powerful contextualized representations.
- We conduct extensive experiments on popular benchmark datasets including SumMe and TVSum. The experimental results clearly demonstrate the effectiveness of the proposed method.

The remaining parts of this paper are organized as follows. Section 2 briefly reviews the related work. Then, the proposed method is described in Section 3. Section 4 shows the experimental results and analysis. Finally, we conclude this work and provide the future prospect in Section 5.

2. Related work

This section briefly reviews the related methods, broadly including two topics: video summarization and attention mechanism, which are discussed in the following.

2.1. Video summarization

Creating high-quality video summaries has remained a continual challenge, prompting researchers to investigate numerous promising

methods over the years. Existing methods can be broadly categorized into two groups: traditional methods, which rely on conventional techniques, and deep learning-based methods which harness the power of neural networks. Traditional methods typically utilize the unsupervised learning paradigm to identify and curate key frames or shots that encapsulate the core content. Cluster-based methods (Chu, Song, & Jaimes, 2015; De Avila et al., 2011; Zhuang, Rui, Huang, & Mehrotra, 1998) represent an initial and straightforward attempt at video summarization. The objective is to cluster visually similar frames or shots and designate their centers as representative summaries of the original content. On the other hand, dictionary learning belongs to the realm of unsupervised methods. It entails the selection of representative elements from the video to construct a summary dictionary capable of accurately reconstructing the content. For example, Mei et al. (2014) put forward a sparse dictionary selection method by $L_{2,0}$ norm. Wang et al. (2016) introduced a similar inhibition constraint for increasing the diversity of summaries. Nevertheless, the performance of these traditional methods remains suboptimal due to their limited representation capability, highlighting the need for deep learning methods.

The fact that videos are displayed frame by frame underscores the importance of aggregating the temporal cues within the video sequence and this core idea has been widely studied in computer vision (Cui, 2022; Zhang, Guo, Wu, Li & Tao, 2023; Zhang, Kang, Liu and Zhu, 2023; Zhang, Zhang, Wu and Tao, 2023). Due to the outstanding modeling capability, RNN-based methods achieve substantial improvement (Fu & Wang, 2021; Zhong, Wang, Zou, Hong, & Hu, 2021). To predict importance scores, Zhang et al. (2016b) utilized bidirectional LSTM to model the forward and backward dependencies, and introduced determinantal point processes (DPP) to increase the diversity of summary content. On the basis of the hierarchical LSTM network (Zhao et al., 2017), Zhao et al. (2018) incorporated shot boundary detection and sequence modeling into one unified method to select key shots. Zhou, Qiao, and Xiang (2018) proposed a reinforcement learning method for video summarization, which comprehensively considered the dissimilarity and representativeness of summary results. Mahasseni, Lam, and Todorovic (2017) combined a LSTM-based key frame selector with a discriminator to generate video summarization through adversarial learning. Apostolidis, Adamantidou, Metsai, Mezaris, and Patras (2021) introduced the Actor-Critic model into the summarization task and tried to learn a policy to select key shots. In practice, these RNN-based networks are generally hampered by their expensive computational cost. To achieve parallel computing, Rochan et al. (2018) employed the fully convolutional sequence model to label each frame.

Although these methods have proved to be effective, they either ignore the pairwise relationships across video frames or the local contextual information within the video sequence, both of which are essential to video understanding. Additionally, in these methods that simultaneously learn global and local contextual information, they might still face the situation of inappropriate video sequence partitioning, which leads to difficulties in fully and effectively modeling contextual information. Our AMFM can overcome the aforementioned issues and successfully outperform state-of-the-art methods.

2.2. Attention mechanism

The attention mechanism, which mimics the selective cognitive function of humans in focusing on relevant information, has emerged as a powerful deep learning technique (Niu, Zhong, & Yu, 2021). As an exceptional method for processing sequential data, the self-attention block initially showed remarkable performance in the machine translation task (Vaswani et al., 2017). It assigns weights to each element by calculating the pairwise relationship, allowing the current location to access all positions without considering their distance. In comparison to sequential models like LSTM, self-attention overcomes the inherent issues of RNNs, such as inefficient computing parallelization and historical information decay with increasing sequence length.

Currently, attention-based methods have exhibited remarkable progress in a wide range of domains, including object detection (Carion et al., 2020), image segmentation (Cheng, Misra, Schwing, Kirillov, & Girdhar, 2022), speech recognition (Yeh et al., 2021), and person re-identification (Zhao, Wang et al., 2022). For instance, Mao, Yang, Lin, Xuan, and Liu (2022) proposed positional attention-guided transformer-like architecture to model features within and across the visual and language modalities. Yang, Miech, Sivic, Laptev, and Schmid (2022) leveraged self-attention to jointly model spatial and visual-linguistic interactions. Badamdorj, Rochan, Wang, and Cheng (2022) proposed a simple contrastive learning method to detect video highlights, which were directly selected by attention scores. For the video summarization task, Ji, Xiong, Pang, and Li (2020) combined RNNs with self-attention to mimic the way of selecting the shots of humans. Fu and Wang (2021) designed a self-attention binary neural tree to address shot-level video summarization, where feature representations are learned from coarse to fine. Zhu et al. (2022) presented a hierarchical attention model via multi-scale features. In particular, it tried to learn the intra-block attention and the inter-block attention, all of which including frame embeddings are immediately concatenated and fed into a scoring module.

Motivated by the success of the attention mechanism, we design an attention-based feature enhancement module to guarantee more effective features by estimating the potential visual relationship across frame-level and video-level embeddings. Moreover, we incorporate attention mechanism and temporal convolution into a unified learnable module to robustly learn multi-granularity contextual information within videos. Finally, we reflect on the significant difference in the semantic scale across global and local temporal cues and adopt a fusion module, which achieves adaptive fusion by learning attention-based fusion weights. Experimental results on standard datasets demonstrate the effectiveness of the proposed method.

3. Proposed method

3.1. Preliminary

This section briefly reviews the multi-head attention mechanism (MHAM) in Transformer (Vaswani et al., 2017) since it plays a crucial role in our overall method. In order to make the method description clearer and avoid excessive redundant descriptions, the part involving MHAM simply uses symbols instead of listing the detailed calculation process.

Concretely, MHAM takes a query matrix $\boldsymbol{Q}$, a key matrix $\boldsymbol{K}$, and a value matrix $\boldsymbol{V}$ as input and maps them to different representation subspaces using linear transformation layers. These features are enriched with global dependencies according to scaled dot-product attention. The final features can be obtained by concatenating all outputs of different subspaces, followed by a linear transformation layer. Mathematically, it is represented as:

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\boldsymbol{H}_1,\boldsymbol{H}_2,\ldots,\boldsymbol{H}_h)W^o \tag{1}$$

where

$$\boldsymbol{H}_i = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{W}_i^q(\boldsymbol{K}\boldsymbol{W}_i^k)^{\mathrm{T}}}{\sqrt{d_h}}\right)\boldsymbol{V}\boldsymbol{W}_i^v \tag{2}$$

where h denotes the number of attention heads, which is simply set to 1 in this paper to save parameters. $\boldsymbol{H}_i$ is the output of ith attention head. d_h is used for scaling. $\boldsymbol{W}_i^q$, $\boldsymbol{W}_i^k$, $\boldsymbol{W}_i^v$, and $\boldsymbol{W}_i^o$ are learnable parameters. MHAM performs self-attention when $\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}$.

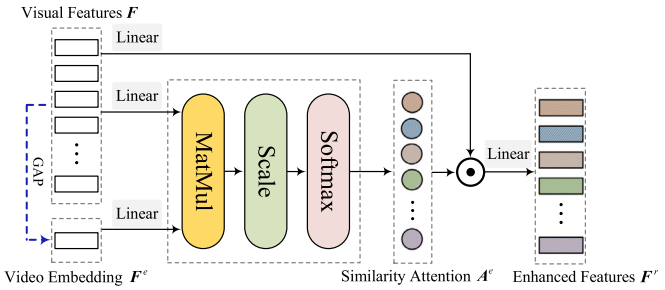

Fig. 2. Pipeline of the CAE module, which primarily aims to learn attention for each frame to enhance visual features by weighting.

3.2. Overview

Fig. 1 illustrates the overview of the proposed method. Specifically, given an input video with N frames, AMFM initially utilizes a pre-trained feature extractor to encode frames, forming visual features $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N]$, where $\boldsymbol{f}_i \in \mathbb{R}^D$ is the feature vector of ith frame and D indicates the dimension. These features are fed into CAE to be enhanced by referring to the potential visual relationship across frame-level and video-level embeddings. Afterward, both coarse-grained and fine-grained contextual information is simultaneously captured by MGE. Next, SAF adaptively fuses multi-granularity contextual information by learning fusion attention. Finally, the fused contextual information is fed into the prediction head to compute importance scores $\boldsymbol{S} = [s_1, s_2, \ldots, s_N]$, which are used to select key shots under a duration constraint.

3.3. Content-aware enhancement module

The majority of existing methods usually exploit visual features extracted by pre-trained deep networks to perform subsequent tasks. However, these independent features only reflect the visual content at the current position and do not consider their interrelationships with the entire video content, which might result in bottlenecks in video understanding. Inspired by the success of the attention mechanism, we propose CAE to address this limitation, which enables each frame to receive informative guidance signals through modeling the similarity relationship across frame-level and video-level representations. By using the idea, frames related to video content are given higher attention, while irrelevant backgrounds are suppressed, further improving the discriminability of features.

The pipeline of CAE is depicted in Fig. 2. Specifically, similar to Xiao et al. (2020), we initiate the process by defining a video-level embedding $\boldsymbol{F}^e \in \mathbb{R}^{1\times D}$. This is achieved through global average pooling (GAP) along the temporal dimension, which allows us to obtain a general representation of video content. Mathematically, the calculation can be written as follows:

$$\boldsymbol{F}^e = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{f}_i \tag{3}$$

Through MHAM, we compute the similarity attention $\boldsymbol{A}^e \in \mathbb{R}^{1\times N}$ by setting $\boldsymbol{F}^e$ and $\boldsymbol{F}$ as the query and key matrices, respectively, which is used to reveal how ith frame is similar to the input video itself. Subsequently, the similarity attention is exploited to form weighted representations $\boldsymbol{F}^{inter} \in \mathbb{R}^{N\times D}$. This calculation can be written as:

$$\boldsymbol{F}^{inter} = \widetilde{\boldsymbol{A}^e}^{\mathrm{T}} \odot \boldsymbol{F}\boldsymbol{W}^e \tag{4}$$

where $\widetilde{\boldsymbol{A}^e} \in \mathbb{R}^{D\times N}$ is obtained by repeating $\boldsymbol{A}^e$. $\odot$ denotes element-wise production. $\boldsymbol{W}^e \in \mathbb{R}^{D\times D}$ is the projection parameter to be learned. The enhanced features $\boldsymbol{F}^r \in \mathbb{R}^{N\times D}$ are calculated by a linear transform. Based on CAE, our method learns discriminative visual features by assigning attention to frames with different video content correlations.

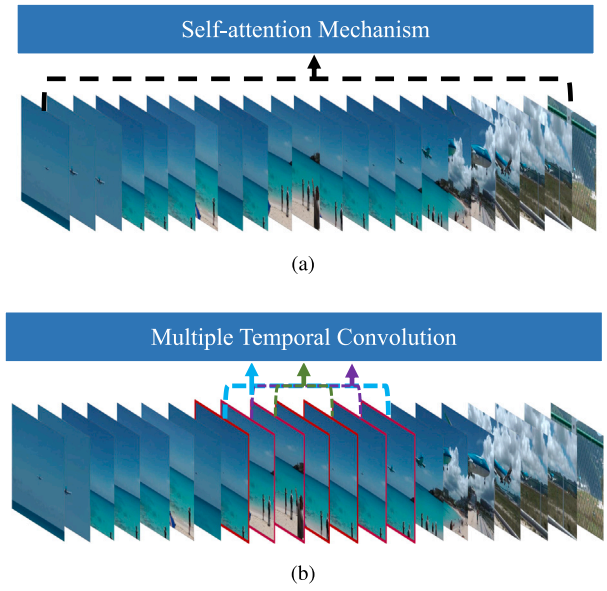

Fig. 3. Illustration of MGE, which consists of (a) a coarse-grained stream and (b) a fine-grained stream to simultaneously model global and local contextual information.

3.4. Multi-granularity encoder

Temporal cues are of great essence to video understanding (Wang et al., 2021). Modeling the entire video sequence is capable of providing a coarse-grained view of the storyline while modeling the local sequences can effectively tackle the detailed information happening in certain period segments. We present MGE, which consists of a coarse-grained stream (CGS) and a fine-grained stream (FGS) to aggregate global and local contextual information based on enhanced visual features. The core idea is shown in Fig. 3.

(1) Coarse-grained Stream: The long-range modeling capability and the computing efficiency can be measured by the maximum path length and the minimum number of sequential operations (Vaswani et al., 2017). The self-attention mechanism shows great advantages in both aspects compared to RNNs. Regarding global contextual information, we employ self-attention to obtain the responses at all positions, which can dramatically reduce the expensive computing cost brought by RNNs and achieve remarkable temporal aggregation. Particularly, CGS begins by projecting the enhanced features $\boldsymbol{F}^r$ into $\boldsymbol{F}^r_q$, $\boldsymbol{F}^r_k$, and $\boldsymbol{F}^r_v$, respectively, followed by pairwise attention calculation and feature aggregation. Simply put, the globally contextualized information $\boldsymbol{G} \in \mathbb{R}^{N\times D}$ can be represented by:

$$\boldsymbol{G} = \mathrm{MultiHead}(\boldsymbol{F}^r, \boldsymbol{F}^r, \boldsymbol{F}^r) \tag{5}$$

(2) Fine-grained Stream: FGS concentrates on modeling temporal cues within tiny local windows to finely understand the video storyline. Specifically, temporal convolution is selected as our fine-grained context aggregator due to its excellent performance in extracting local features in computer vision (Ke, Sun, Li, Yan, & Lau, 2022; Liang, Guo, Li, Jin and Shen, 2022; Zhang, Song and Li, 2023). Multi-granularity features can provide more valuable assistance, which encourages us to specify more than one window size. Relied on the consideration of summarization effectiveness and the number of parameters, the candidate sizes are pre-denoted as $\{k_1 = 3, k_2 = 5, k_3 = 7\}$, respectively. Accordingly, to alleviate the computational burden arising from multiple windows, we further exploit depth-separable convolution to achieve our purpose. Technically, the fine-grained stream starts with encoded representation $\boldsymbol{F}^r_v$ to model representations from the same latent space

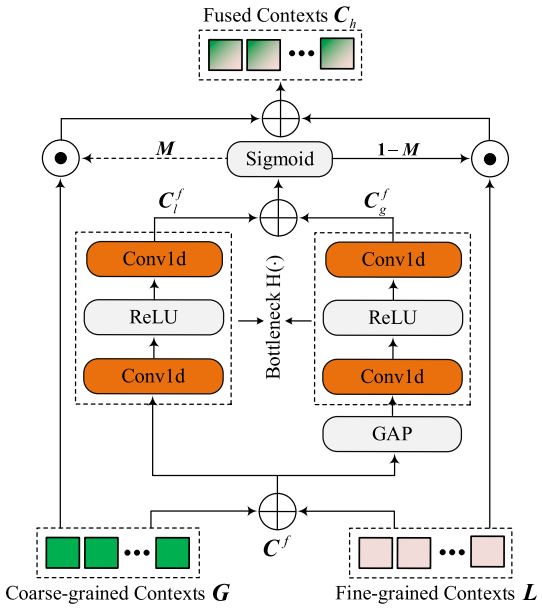

Fig. 4. Illustration of the SAF module, which consists of dual pathways to adaptively perform deep context fusion by learning fusion attention.

with the coarse-grained stream. The result regarding window k_i ($i = 1, 2, 3$) can be calculated by:

$$\boldsymbol{O}^{k_i} = \text{PConv}(\text{DConv}_{k_i}(\boldsymbol{F}_v^r)) \tag{6}$$

where PConv(·) and DConv(·) denote the pointwise and depthwise convolution. Since they are aggregated within extremely similar frame ranges, we form the final fine-grained contextual information $\boldsymbol{L} \in \mathbb{R}^{N\times D}$ by directly summing them up:

$$\boldsymbol{L} = \sum_{i=1}^{w} a_i \boldsymbol{O}^{k_i} \tag{7}$$

where $w = 3$ and $a_i \in \{0, 1\}$ denotes whether window k_i participates in calculation. Our default model only incorporates k_1 and k_2, which will be discussed in Section 4.

3.5. Scale-adaptive fusion module

The purpose of context fusion is to gather the positive aspects of features across multiple levels to represent video content in a condensed form. Typically, intuitive fusion strategies (*e.g.,* summation) are often preferred due to their ease of computation. However, coarse-grained and fine-grained contextual information cover a significant difference in the semantic scale, hence, these simple operations cannot achieve sufficient context fusion owing to limited information exchange. To address this limitation, we propose SAF to adaptively perform more effective fusion across coarse-grained and fine-grained contextual information learned by MGE by learning fusion attention.

As depicted in Fig. 4, SAF is composed of dual-learning pathways to aggregate features comprehensively. Similar to the bottleneck layer in ResNet (He, Zhang, Ren, & Sun, 2016), each pathway includes a bottleneck structure H(·) that consists of two temporal convolution layers and an activation function sandwiched by them, where the first temporal convolution reduces the feature dimension from D to D/m, followed by another temporal convolution to restore the dimension to D. Concretely, the first pathway is responsible for learning local attention $\boldsymbol{C}_l^f \in \mathbb{R}^{N\times D}$ from initial fusion features $\boldsymbol{C}^f \in \mathbb{R}^{N\times D}$ obtained by summation. The second pathway performs GAP before the bottleneck

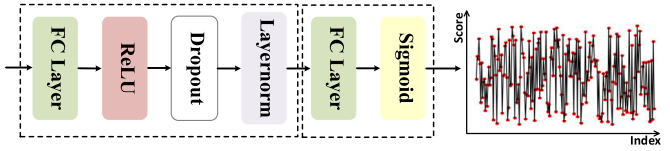

Fig. 5. Pipeline of the prediction head, which mainly consists of two fully connected layers and a sigmoid function for importance score prediction.

layer to model global attention $\boldsymbol{C}_g^f \in \mathbb{R}^{1\times D}$. Formally, this can be formulated as follows:

$$\boldsymbol{C}_l^f = \text{H}(\boldsymbol{C}^f) \tag{8}$$

$$\boldsymbol{C}_g^f = \text{H}(\delta(\boldsymbol{C}^f)) \tag{9}$$

where $\delta(\cdot)$ denote the GAP operation. Subsequently, we incorporate the learned attentions into a unified representation form, which is followed by a sigmoid function to generate an adaptive attention matrix $\boldsymbol{M} \in \mathbb{R}^{N\times D}$:

$$\boldsymbol{M} = \sigma\left(\boldsymbol{C}_l^f + \boldsymbol{C}_g^f\right) \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function. Based on the attention matrix, the fused contextual representations $\boldsymbol{C}_h \in \mathbb{R}^{N\times D}$ are formed by a weighted averaging. Mathematically, this process can be written as:

$$\boldsymbol{C}_h = \boldsymbol{M} \odot \boldsymbol{G} + (\boldsymbol{1} - \boldsymbol{M}) \odot \boldsymbol{L} \tag{11}$$

Leveraging this module, our method is capable of conducting comprehensive and adaptive context fusion while preserving important information to accurately represent the storyline. Moreover, the well-designed bottleneck structure significantly reduces the training burden, resulting in a concise and feasible method.

3.6. Summary generation

The proposed method includes a prediction head, in addition to the components mentioned before, which is responsible for predicting importance scores for each frame according to the fused contextual information $\boldsymbol{C}_h$. Its architecture, depicted in Fig. 5, consists primarily of two fully connected layers. The final summary of an input video is created by selecting a set of sub-shots. In particular, following previous efforts, we exploit the Kernel Temporal Segmentation (KTS) (Potapov et al., 2014) to detect change points, which are used for segmenting an entire video into U disjoint subsequences. Then, we convert frame-level importance scores to shot-level importance scores by taking an average within each shot:

$$p_i = \frac{1}{l_i} \sum_{k=1}^{l_i} s_k \tag{12}$$

where p_i and l_i are importance score and duration of ith shot, respectively. The duration of a summary is limited to no more than 15% of the total duration. Next, we formulate the selection of a key-shot-based summary as a knapsack problem, which can be mathematically represented as:

$$\max \sum_{i=1}^{U} x_i p_i, \quad s.t. \sum_{i=1}^{U} x_i l_i \le 0.15 \times N \tag{13}$$

where $x_i \in \{0, 1\}$ indicates whether ith shot is selected or not. We solve this problem by dynamic programming and generate summarization by concatenating those shots with $a_i = 1$ in chronological order.

3.7. Network training

To measure the difference between the ground truth scores and predicted scores, we adopt the mean-square error (MSE) as our objective function $\mathcal{L}$ to iteratively optimize the model parameters during the training process. The form can be written as:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(s_i - y_i)^2 \tag{14}$$

where θ is the parameters of our AMFM. y_i is the human-created importance score of ith frame.

4. Experiment

This section focuses on evaluating the summarization performance of our method on benchmark datasets. We start by providing the experimental setup, which includes the description of datasets, evaluation metrics, and implementation details. Then, we present quantitative results, ablation study, and qualitative results.

4.1. Experimental setup

(1) Datasets: We evaluate our method using two well-established benchmark datasets including SumMe (Gygli et al., 2014) and TVSum (Song, Vallmitjana, Stent, & Jaimes, 2015). The SumMe dataset consists of 25 videos depicting various events, such as food and sports, and each video is annotated by at least 15 users. The duration of each video ranges from 1 to 6 min. The TVSum dataset comprises 50 videos from 10 categories, with each video annotated by 20 users. The duration of the videos varies from 2 to 10 min. In addition, we employ YouTube (De Avila et al., 2011) and OVP (De Avila et al., 2011) to augment our training data, both of which are created with key frame-level labels and contain 89 videos in total. The datasets used in our experiments contain videos with quick and slow scene changes, posing challenges in evaluating the effectiveness of the proposed method. Regarding the evaluation setting, following previous methods, 80% videos of each dataset are selected for training and 20% for testing. Existing methods primarily adopt 5 random splits (5 Random), 10 random splits (10 Random), and multiple random splits (M Random) and report average summarization performance. Nevertheless, such split manners inevitably can cause certain videos in the dataset to be used or omitted multiple times, leading to unfair performance comparisons due to inappropriate data splitting schema. To alleviate this problem, this paper follows He et al. (2019), Jung et al. (2019), Li, Ke, Gong and Zhang (2023), Liang, Lv et al. (2022) and Zhou et al. (2018), adopting standard 5-fold cross-validation (5 FCV) to ensure that each video participates in testing procedure, and thus generates more reliable experimental results.

(2) Evaluation Metric: In order to evaluate our method comprehensively with other state-of-the-art methods, we first report the F-score performance on the SumMe and TVSum datasets. Suppose $\boldsymbol{X}_p$ and $\boldsymbol{X}_h$ represent the predicted summaries and human summaries, respectively. The F-score can be computed by:

$$P = \frac{\text{overlapped duration of } \boldsymbol{X}_p \text{ and } \boldsymbol{X}_h}{\text{duration of } \boldsymbol{X}_p} \tag{15}$$

$$R = \frac{\text{overlapped duration of } \boldsymbol{X}_p \text{ and } \boldsymbol{X}_h}{\text{duration of } \boldsymbol{X}_h} \tag{16}$$

$$F\text{-}score = \frac{2\times P\times R}{P+R}\times 100\% \tag{17}$$

Moreover, it has been studied in a recent study (Otani, Nakashima, Rahtu, & Heikkila, 2019) that the F-score metric may not be sensitive enough to the differences in the importance score computation, thus we also use Kendall's τ and Spearman's ρ to calculate correlation coefficients between the ground truth scores and predicted scores.

(3) Implementation Detail: We employ GoogLeNet (Szegedy et al., 2015) pre-trained on the ImageNet (Deng et al., 2009) to extract visual features. The output of the pool-5 layer is taken as the feature vector to represent the visual content. It indicates that the dropout, fully connected layer, and softmax are excluded. The dimension D of visual features is 1024. The number of attention heads is set to 1 for saving parameters. By default, we use convolution kernels of size 3 and 5 for fine-grained context modeling in MGE. In SAF, by sensitivity analysis, we set the reduction rate $m = 4$. Moreover, we adopt convolution kernels of size 3 and 1 for feature learning in the first and second pathways, respectively. We train our model using the Adam optimizer with the learning rate of 2×10^{-5} for the SumMe dataset, 5×10^{-5} for the TVSum dataset. The training process is terminated after 300 epochs.

4.2. Comparisons with state-of-the-art methods

(1) Baselines: We perform comprehensive experiments to compare the effectiveness of the proposed method with that of existing state-of-the-art methods, across different evaluation manners. Specifically, these methods can be categorized into traditional methods including TVSum (Song et al., 2015), DPP (Zhang, Chao, Sha, & Grauman, 2016a), ERSUM (Li, Zhao, & Lu, 2017), MSDS-CC (Meng, Wang, Wang, Yuan, & Tan, 2017), and deep learning-based methods including vsLSTM (Zhang et al., 2016b), dppLSTM (Zhang et al., 2016b), SUM-GAN (Mahasseni et al., 2017), DR-DSN (Zhou et al., 2018), FCSN (Rochan et al., 2018), SASUM (Wei et al., 2018), HSA-RNN (Zhao et al., 2018), ACGAN (He et al., 2019), CSNet (Jung et al., 2019), A-AVS (Ji et al., 2020), M-AVS (Ji et al., 2020), RSGN (Zhao, Li, Lu and Li, 2022), DHAVS (Lin, Zhong, & Fares, 2022), LMHA (Zhu et al., 2022), CAAN (Liang, Lv et al., 2022), HMT (Zhao, Gong et al., 2022), VJMHT (Li, Ke, Gong and Zhang, 2023), and SSPVS (Li, Ke, Gong and Drummond, 2023).

(2) Comparisons Under the Standard Setting: Table 1 presents the experimental results for the standard setting on the SumMe and TVSum datasets. The results demonstrate that the proposed method performs exceptionally well in comparison to existing state-of-the-art methods. Notably, the values reported indicate that the deep learning-based methods, which leverage feature representations with rich semantic information and effective modeling of temporal cues, generally outperform traditional methods. AMFM is based solely on matrix operations and does not use any recursive structures used in works like (Yuan et al., 2019; Zhang et al., 2016b; Zhao, Li et al., 2022). This design choice enables efficient GPU parallelization. Although our method achieves comparable performance to M-AVS (Ji et al., 2020) and LHMA (Zhu et al., 2022) on the TVSum dataset, all the experiments are conducted using standard cross-validation instead of a random split, indicating comprehensive and reliable evaluation results as discussed in Section 4.1. Compared with those efforts adopting the 5 FCV test method, AMFM surpasses them by a large margin in the F-score evaluation, which can be attributed to the excellent extraction and processing capability of our meticulous model. By observing the provided parameters, our method can achieve a good trade-off compared with other state-of-the-art methods.

(3) Comparisons Under the Augmented and Transfer Settings: We further compare the proposed method with state-of-the-art methods under the augmented and transfer settings, as shown in Table 2. We can observe that AMFM achieves an encouraging performance. The transfer setting is an effective but challenging method to verify the transferability of the model. The reported performance values on both datasets illustrate that AMFM can learn meaningful semantic information from videos from other domains. The performance under the transfer setting is significantly lower, indicating that cross-dataset learning remains a difficult problem that requires further investigation.

(4) Comparisons of Rank Correlation Coefficients: Moreover, we also take into account the rank order statistics on the TVSum dataset to evaluate the effectiveness of different methods, as recommended by Otani et al. (2019). Table 3 presents the results, which indicate that random

Table 1
Comparisons of the F-score (%) and the number of parameters (M) with state-of-the-art methods under the standard evaluation setting.

Method	Shot segmentation	Feature	SumMe ↑	TVSum ↑	Params ↓	Test method
TVSum (Song et al., 2015)	Change-point detection	HoG+GIST+SIFT	–	50.0	–	–
DPP (Zhang et al., 2016a)	KTS	AlexNet	40.9	–	–	5 Random
ERSUM (Li et al., 2017)	Uniform segmentation	VGGNet-16	43.1	59.4	–	–
MSDS-CC (Meng et al., 2017)	KTS	GIST+GoogLeNet	40.6	52.3	–	–
vsLSTM (Zhang et al., 2016b)	KTS	GoogLeNet	37.6	54.2	2.63	5 Random
dppLSTM (Zhang et al., 2016b)	KTS	GoogLeNet	38.6	54.7	2.63	5 Random
SUM-GAN (Mahasseni et al., 2017)	KTS	GoogLeNet	41.7	56.3	295.86	5 Random
FCSN (Rochan et al., 2018)	KTS	GoogLeNet	47.5	56.8	36.58	M Random
SASUM (Wei et al., 2018)	KTS	InceptionV3	45.3	58.2	44.07	10 Random
HSA-RNN (Zhao et al., 2018)	Change-point detection	VGGNet-16	42.3	58.7	4.20	5 Random
A-AVS (Ji et al., 2020)	KTS	GoogLeNet	43.9	59.4	4.40	5 Random
M-AVS (Ji et al., 2020)	KTS	GoogLeNet	44.4	61.0	4.40	5 Random
RSGN (Zhao, Li et al., 2022)	KTS	GoogLeNet	45.0	60.1	–	5 Random
DHAVS (Lin et al., 2022)	KTS	3D ResNeXt-101	45.6	60.8	–	5 Random
LMHA (Zhu et al., 2022)	KTS	GoogLeNet	51.1	61.0	–	5 Random
HMT (Zhao, Gong et al., 2022)	KTS	GoogLeNet	44.1	60.1	–	5 Random
DR-DSN (Zhou et al., 2018)	KTS	GoogLeNet	42.1	58.1	2.63	5 FCV
ACGAN (He et al., 2019)	KTS	GoogLeNet	47.2	59.4	–	5 FCV
CSNet (Jung et al., 2019)	KTS	GoogLeNet	48.6	58.5	–	5 FCV
CAAN (Liang, Lv et al., 2022)	KTS	GoogLeNet	50.6	59.3	–	5 FCV
VJMHT (Li, Ke, Gong and Zhang, 2023)	KTS	GoogLeNet	50.6	60.9	35.44	5 FCV
SSPVS (Li, Ke, Gong and Drummond, 2023)	KTS	GoogLeNet	48.7	60.3	–	5 FCV
AMFM	KTS	GoogLeNet	51.8	61.0	13.66	5 FCV

Table 2
Comparisons of the F-score (%) with state-of-the-art methods under the canonical (C), augmented (A), and transfer (T) settings, respectively.

Method	SumMe ↑			TVSum ↑		
	C	A	T	C	A	T
vsLSTM (Zhang et al., 2016b)	37.6	41.6	40.7	54.2	57.9	56.9
dppLSTM (Zhang et al., 2016b)	38.6	42.9	41.8	54.7	59.6	58.7
SUM-GAN (Mahasseni et al., 2017)	41.7	43.6	–	56.3	61.2	–
DR-DSN (Zhou et al., 2018)	42.1	43.9	42.6	58.1	59.8	58.9
FCSN (Rochan et al., 2018)	47.5	51.1	44.1	56.8	59.2	58.2
HSA-RNN (Zhao et al., 2018)	42.3	42.1	–	58.7	59.8	–
CSNet (Jung et al., 2019)	48.6	48.7	44.1	58.5	57.1	57.4
A-AVS (Ji et al., 2020)	43.9	44.6	–	59.4	60.8	–
M-AVS (Ji et al., 2020)	44.4	46.1	–	61.0	61.8	–
RSGN (Zhao, Li et al., 2022)	45.0	45.7	44.0	60.1	61.1	60.0
DHAVS (Lin et al., 2022)	45.6	46.5	43.5	60.8	61.2	57.5
LMHA (Zhu et al., 2022)	51.1	52.1	45.4	61.0	61.5	55.1
HMT (Zhao, Gong et al., 2022)	44.1	44.8	–	60.1	60.3	–
VJMHT (Li, Ke, Gong and Zhang, 2023)	50.6	51.7	46.4	60.9	61.9	58.9
SSPVS (Li, Ke, Gong and Drummond, 2023)	48.7	50.4	45.8	60.3	61.8	57.8
AMFM	51.8	52.8	46.4	61.0	60.8	58.6

Table 3
Comparisons of rank correlation coefficients with state-of-the-art methods.

Method	Kendall's τ ↑	Spearman's ρ ↑
Random (Otani et al., 2019)	0.000	0.000
Human (Otani et al., 2019)	0.177	0.204
dppLSTM (Zhang et al., 2016b)	0.042	0.055
DR-DSN (Zhou et al., 2018)	0.020	0.026
HSA-RNN (Zhao et al., 2018)	0.082	0.088
CAAN (Liang, Lv et al., 2022)	0.038	0.050
DHAVS (Lin et al., 2022)	0.082	0.089
RSGN (Zhao, Li et al., 2022)	0.083	0.090
HTM (Zhao, Gong et al., 2022)	0.096	0.107
VJMHT (Li, Ke, Gong and Zhang, 2023)	0.097	0.105
SSPVS (Li, Ke, Gong and Drummond, 2023)	0.177	0.233
AMFM	0.179	0.233

Table 4
Comparisons of elapsed time (millisecond).

Dataset (Average duration)	SumMe ↓ (2min 26s)	TVSum ↓ (3min 55s)
DR-DSN (Zhou et al., 2018)	5.18	7.87
AMFM	5.24	5.61

summary performs the worst and is entirely irrelevant to manual annotations. Compared to existing state-of-the-art methods under this more reliable evaluation metric, our method yields significantly superior performance and even surpasses human summaries. This finding suggests that there may be internal inconsistency within human annotations, whereas our method can effectively capture multiple users' preferences. Besides, our method can accurately capture the storyline by robustly mining and fusing global and local information, enabling it to precisely predict the importance of each frame. It is noteworthy that SSPVS yields comparable performance with AMFM in these correlation coefficients, possibly owing to it being trained on an extremely larger-scale dataset. However, our AMFM, trained on a small dataset, achieves comparable or even superior performance, thus demonstrating the effectiveness of the proposed method.

(5) Comparisons of Elapsed Time: To investigate the inference efficiency of AMFM, we provide experimental results about our method and a RNN-based method in Table 4. Here, as a classic and effective architecture, DR-DSN (Zhou et al., 2018) is chosen as the target for comparison, which only includes a concise Bi-LSTM structure. Since our method is obviously more complex, the reported values can highlight our advantage in terms of computing time. All the results are conducted under a completely fair experimental environment using a 2080Ti GPU. After obtaining all inference times of videos on the SumMe and TVSum

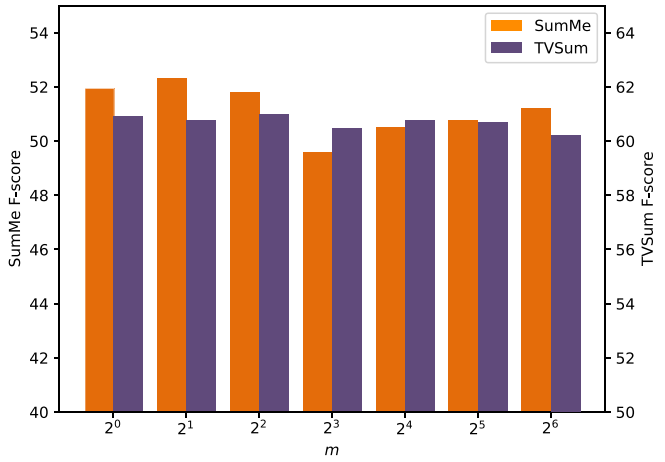

Fig. 6. Ablation study on the dimension reduction rate m.

Table 5
Ablation study on dominant components. ✓ denotes whether the corresponding module is included in the model.

Exp.	Component settings					SumMe ↑	TVSum ↑
	CAE	MGE	Sum.	Concat.	SAF		
1						47.2	56.7
2	✓					48.7	58.9
3	✓	✓	✓			50.1	60.7
4	✓	✓		✓		50.7	59.5
5	✓	✓			✓	51.8	61.0

datasets, we report the average values, respectively. In addition, to eliminate the time deviation caused by irrelevant factors, we only consider the time it takes to predict the importance score, excluding pre-processing and post-processing. It can be found that AMFM has significantly less running time on the TVSum datasets and a comparable running time on the SumMe dataset compared with DR-DSN. This is because the TVSum dataset has a longer duration, hence, recursive neural networks need to execute recursively more times while our architecture can support parallel computing.

4.3. Ablation study

There are three dominant components in AMFM, including CAE, MGE, and SAF. To verify the necessity and contribution of each component, we implement ablation experiments and conduct analysis.

(1) Study on Dimension Reduction Rate: We conduct a thorough analysis of parameter m in SAF to examine its effect on information transmission. To obtain the most suitable parameter for video summarization, we sequentially set m to 1 to 64 and report the experimental results on the SumMe dataset and TVSum dataset. The results are presented in Fig. 6. We can observe that the F-scores are lower as m increases and it can be explained by that significant dimension reduction would lead to more information loss. To balance the summarization performance between the SumMe and TVSum datasets, we set m to 4 as our default value, which determines the intermediate feature dimension in the SAF module.

(2) Study on Dominant Components: The contribution of different components has been thoroughly examined, and the results are summarized in Table 5. In our initial experiment (Exp. 1), the base model only includes the prediction head, directly predicting importance scores based on the pre-trained visual features. To enhance the representation capability of these visual features, we introduce CAE to the base model (Exp. 2), which leads to significant improvements of 1.5% and 2.2% in F-scores on the SumMe and TVSum datasets, respectively.

Table 6
Ablation study on multi-granularity information. ✓ and ✗ denote the corresponding contextual information is learned or not.

Exp.	Granularity settings		SumMe ↑	TVSum ↑
	Coarse-grained	Fine-grained		
1	✗	✗	48.7	58.9
2	✓	✗	48.8	60.0
3	✗	✓	50.0	60.1
4	✓	✓	51.8	61.0

Table 7
Ablation study on local window. ✓ denotes whether the corresponding temporal convolution kernel is used in the MGE module.

Exp.	Local window settings			SumMe ↑	TVSum ↑
	$k_1 = 3$	$k_2 = 5$	$k_3 = 7$		
1	✓			50.6	60.4
2		✓		51.4	60.4
3			✓	51.9	60.2
4	✓	✓		51.8	61.0
5	✓		✓	51.9	60.4
6		✓	✓	53.0	60.3
7	✓	✓	✓	51.9	60.6

Subsequently, we further integrate MGE into our architecture to model contextual information within videos. By obtaining both coarse-grained and fine-grained contextual information, we implement different fusion strategies, including summation (Sum.), concatenation (Concat.), and SAF. The summation-based model directly fuses multi-granularity contexts by element-wise summation and the concatenation-based model leverages a fully connected layer to reduce the dimension of concatenated multi-granularity contexts from $2 \times D$ to D. Notably, SAF outperforms other fusion methods by a substantial margin, thanks to its unique adaptive fusion mechanism. In summary, these experimental findings strongly underscore the importance of each individual component in our method.

(3) Study on Multi-Granularity Information: We explore the impact of coarse-grained and fine-grained contextual information by sequentially removing CGS and FGS in the MGE module. As depicted in Table 6, the model (Exp. 1) devoid of any contextual information exhibits the worst performance across the listed situations, underscoring the pivotal role of context in comprehending video content. Introducing single-granularity contextual information (Exp. 2 and Exp. 3) yields substantial performance enhancements, enhancing the model's performance by a minimum of 1.6% and 3.3% on the SumMe and TVSum datasets, respectively. Notably, fine-grained context confers a more significant performance boost for summarization. This is attributable to our well-designed architecture, which adeptly captures precise content understanding by robustly learning rich temporal cues within small local windows. These experimental results given by our default method (Exp. 4) substantiate the rationality behind our method.

(4) Study on Local Window: To investigate the influence of different local window sizes, we conduct ablation experiments by using different combinations of temporal convolution in FGS of MGE. The experimental results are presented in Table 7. Our method achieves the best performance on the SumMe dataset when utilizing both k_2 and k_3, while on the TVSum dataset, it performs best with k_1 and k_2. These results can be attributed to the nature of the videos in the TVSum dataset, which frequently feature abrupt shot changes. Consequently, employing larger local windows in such cases may result in a less coherent information representation. Furthermore, it is worth noting that the F-scores on the SumMe dataset exhibit significant fluctuations with an absolute difference of 2.4%. This can be attributed to the specific evaluation protocol adopted for this dataset. Considering the summarization performance on both datasets and the parameters, we opt to use k_1 and k_2 as the default configuration to ensure comparable performance with other state-of-the-art methods.

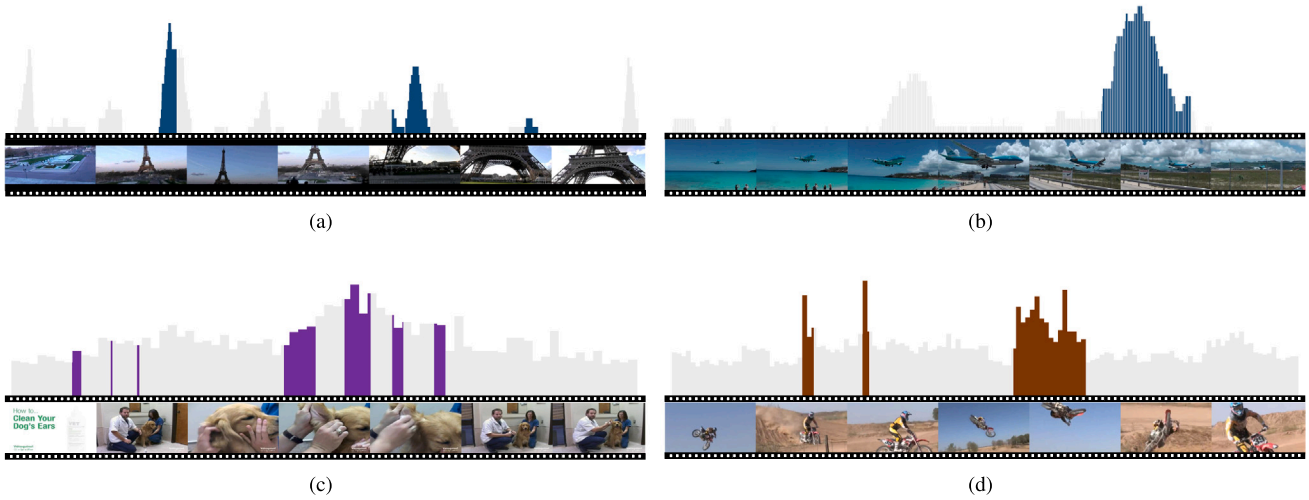

Fig. 7. Qualitative results on the SumMe and TVSum datasets. Four videos are selected as examples, including 9th and 19th videos in SumMe, 15th and 41th videos in TVSum. The x-axis is the frame index. The images below are some example frames in generated summaries by our method.

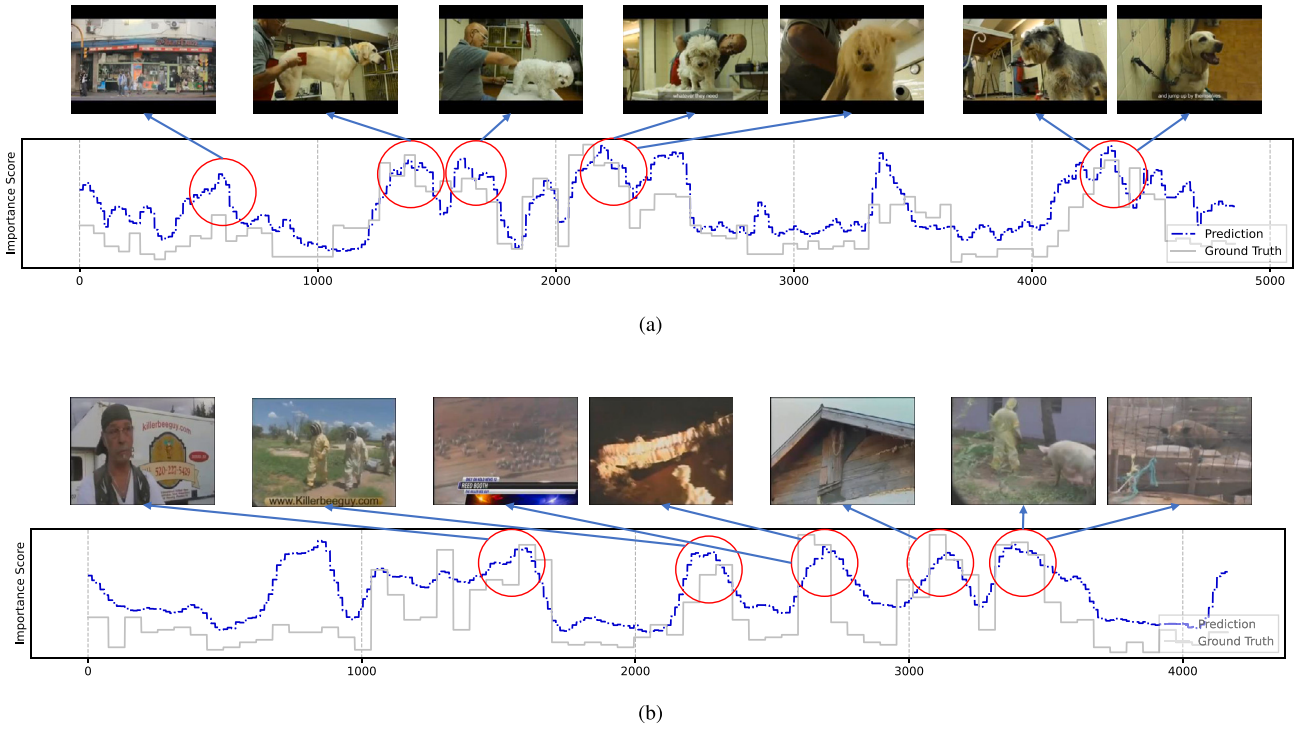

Fig. 8. Correlation visualization results of 14th and 39th video in TVSum. The predicted scores are consistent with manual annotations, indicating that the AMFM can be aware of video content and select important and valuable segments.

4.4. Qualitative results

To visually demonstrate the effectiveness of our method, we present qualitative results in Fig. 7. In this illustration, the light gray bars represent the ground truth summaries, while the colored bars depict the predicted summaries of our method. Additionally, at the bottom of the histogram, we showcase example frames chosen by our method. To ensure a comprehensive evaluation of the summarization performance, we select four videos from both the SumMe and TVSum datasets, encompassing various topics such as landscapes, pets, and sports. These selections allow us to thoroughly validate the performance of our method.

The results affirm that our method excels at identifying the most significant shots, as indicated by their high-importance scores. This underscores the capability of AMFM to effectively capture the primary content of the video. Furthermore, the generated summaries offer a comprehensive narrative of the entire story and encompass a diverse range of content. This diversity enables viewers to quickly grasp the activities depicted in the videos. Additionally, we provide correlation results between ground truth scores and predicted scores in Fig. 8. The results clearly highlight the effectiveness of our method in modeling the relative importance of videos.

Interestingly, our visualization results reveal that despite achieving a high F-score, some peak points are not selected. This phenomenon may be attributed to the inherent limitations of the widely employed

KTS segmentation method, which often prioritizes computational speed at the expense of accuracy. Consequently, it becomes imperative for future research endeavors to concentrate on the development of more advanced shot boundary detection methods that can enhance the performance of video summarization methods.

5. Conclusion

In this work, we propose an effective Attention-Guided Multi-Granularity Fusion Model, which comprises the CAE module, the MGE module, and the SAF module, to sufficiently and effectively model contextual information for video summarization. Specifically, CAE targets to complete the feature enhancement for pre-trained visual features. Then, MGE is introduced to robustly model coarse-grained and fine-grained contextual information. Finally, SAF is used to facilitate adaptive fusion across multi-granularity contextual information with a significant difference in the semantic scale. The proposed AMFM is extensively evaluated on benchmark datasets and demonstrates significantly competitive performance. Moving forward, the integration of multimodal information to generate high-quality summary results that meet human preferences is an interesting avenue for future research.

CRediT authorship contribution statement

Yunzuo Zhang: Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing. **Yameng Liu:** Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Cunyu Wu:** Conceptualization, Investigation, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China (No. 61702347, No. 62027801), the Natural Science Foundation of Hebei Province, China (No. F2022210007, No. F2017210161), the Science and Technology Project of Hebei Education Department, China (No. ZD2022100, No. QN2017132), the Central Guidance on Local Science and Technology Development Fund, China (No. 226Z0501G).

References

Apostolidis, Evlampios, Adamantidou, Eleni, Metsai, Alexandros I., Mezaris, Vasileios, & Patras, Ioannis (2021). AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(8), 3278–3292.

Badamdorj, Taivanbat, Rochan, Mrigank, Wang, Yang, & Cheng, Li (2022). Contrastive learning for unsupervised video highlight detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 14042–14052).

Bettadapura, Vinay, Pantofaru, Caroline, & Essa, Irfan (2016). Leveraging contextual cues for generating basketball highlights. In *Proceedings of the ACM international conference on multimedia* (pp. 908–917).

Carion, Nicolas, Massa, Francisco, Synnaeve, Gabriel, Usunier, Nicolas, Kirillov, Alexander, & Zagoruyko, Sergey (2020). End-to-end object detection with transformers. In *Proceedings of the European conference on computer vision* (pp. 213–229).

Cheng, Bowen, Misra, Ishan, Schwing, Alexander G, Kirillov, Alexander, & Girdhar, Rohit (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1290–1299).

Chu, Wen-Sheng, Song, Yale, & Jaimes, Alejandro (2015). Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3584–3592).

Cui, Yiming (2022). Dynamic feature aggregation for efficient video object detection. In *Proceedings of the Asian conference on computer vision* (pp. 944–960).

De Avila, Sandra Eliza Fontes, Lopes, Ana Paula Brandao, da Luz Jr, Antonio, & de Albuquerque Araújo, Arnaldo (2011). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, *32*(1), 56–68.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, & Fei-Fei, Li (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Fu, Hao, & Wang, Hongxing (2021). Self-attention binary neural tree for video summarization. *Pattern Recognition Letters*, *143*, 19–26.

Gygli, Michael, Grabner, Helmut, Riemenschneider, Hayko, & Gool, Luc Van (2014). Creating summaries from user videos. In *Proceedings of the European conference on computer vision* (pp. 505–520).

He, Xufeng, Hua, Yang, Song, Tao, Zhang, Zongpu, Xue, Zhengui, Ma, Ruhui, et al. (2019). Unsupervised video summarization with attentive conditional generative adversarial networks. In *Proceedings of the ACM international conference on multimedia* (pp. 2296–2304).

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huang, Cheng, & Wang, Hongmei (2020). A novel key-frames selection framework for comprehensive video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(2), 577–589.

Hussain, Tanveer, Muhammad, Khan, Ding, Weiping, Lloret, Jaime, Baik, Sung Wook, & de Albuquerque, Victor Hugo C. (2021). A comprehensive survey of multi-view video summarization. *Pattern Recognition*, *109*, Article 107567.

James, Hale More than 500 hours of content are now being uploaded to YouTube every minute. https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/.

Ji, Zhong, Xiong, Kailin, Pang, Yanwei, & Li, Xuelong (2020). Video summarization with Attention-Based Encoder–Decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(6), 1709–1717.

Jiang, Hao, & Mu, Yadong (2022). Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 16388–16398).

Jung, Yunjae, Cho, Donghyeon, Kim, Dahun, Woo, Sanghyun, & Kweon, In So (2019). Discriminative feature learning for unsupervised video summarization. *Vol. 33*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8537–8544).

Ke, Zhanghan, Sun, Jiayu, Li, Kaican, Yan, Qiong, & Lau, Rynson WH (2022). Modnet: Real-time trimap-free portrait matting via objective decomposition. *Vol. 36*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1140–1147).

Li, Haopeng, Ke, Qiuhong, Gong, Mingming, & Drummond, Tom (2023). Progressive video summarization via multimodal self-supervised learning. In *Proceedings of the IEEE winter conference on applications of computer vision* (pp. 5584–5593).

Li, Haopeng, Ke, Qiuhong, Gong, Mingming, & Zhang, Rui (2023). Video joint modelling based on hierarchical transformer for co-summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3904–3917.

Li, Wenxu, Pan, Gang, Wang, Chen, Xing, Zhen, & Han, Zhenjun (2022). From coarse to fine: Hierarchical structure-aware video summarization. *ACM Transactions on Multimedia Computing Communications and Applications*, *18*(1s).

Li, Xuelong, Zhao, Bin, & Lu, Xiaoqiang (2017). A general framework for edited video and raw video summarization. *IEEE Transactions on Image Processing*, *26*(8), 3652–3664.

Liang, Zhiyuan, Guo, Kan, Li, Xiaobo, Jin, Xiaogang, & Shen, Jianbing (2022). Person foreground segmentation by learning multi-domain networks. *IEEE Transactions on Image Processing*, *31*, 585–597.

Liang, Guoqiang, Lv, Yanbing, Li, Shucheng, Zhang, Shizhou, & Zhang, Yanning (2022). Video summarization with a convolutional attentive adversarial network. *Pattern Recognition*, *131*, Article 108840.

Lin, Tianwei, Zhao, Xu, Su, Haisheng, Wang, Chongjing, & Yang, Ming (2018). Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Lin, Jingxu, Zhong, Sheng-hua, & Fares, Ahmed (2022). Deep hierarchical LSTM networks with attention for video summarization. *Computers & Electrical Engineering*, *97*, Article 107618.

Liu, Tianrui, Meng, Qingjie, Huang, Jun-Jie, Vlontzos, Athanasios, Rueckert, Daniel, & Kainz, Bernhard (2022). Video summarization through reinforcement learning with a 3D spatio-temporal U-Net. *IEEE Transactions on Image Processing*, *31*, 1573–1586.

Mahasseni, Behrooz, Lam, Michael, & Todorovic, Sinisa (2017). Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 202–211).

Mao, Aihua, Yang, Zhi, Lin, Ken, Xuan, Jun, & Liu, Yong-Jin (2022). Positional attention guided transformer-like architecture for visual question answering. *IEEE Transactions on Multimedia*, 1–13.

Mei, Shaohui, Guan, Genliang, Wang, Zhiyong, He, Mingyi, Hua, Xian-Sheng, & Dagan Feng, David (2014). L2,0 constrained sparse dictionary selection for video summarization. In *Proceedings of the IEEE international conference on multimedia and expo* (pp. 1–6).

Meng, Jingjing, Wang, Suchen, Wang, Hongxing, Yuan, Junsong, & Tan, Yap-Peng (2017). Video summarization via multi-view representative selection. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 1189–1198).

Merler, Michele, Mac, Khoi-Nguyen C., Joshi, Dhiraj, Nguyen, Quoc-Bao, Hammer, Stephen, Kent, John, et al. (2019). Automatic curation of sports highlights using multimodal excitement features. *IEEE Transactions on Multimedia*, *21*(5), 1147–1160.

Niu, Zhaoyang, Zhong, Guoqiang, & Yu, Hui (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62.

Otani, Mayu, Nakashima, Yuta, Rahtu, Esa, & Heikkila, Janne (2019). Rethinking the evaluation of video summaries. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7596–7604).

Park, Jungin, Lee, Jiyoung, Kim, Ig-Jae, & Sohn, Kwanghoon (2020). Sumgraph: Video summarization via recursive graph modeling. In *Proceedings of the European conference on computer vision* (pp. 647–663).

Potapov, Danila, Douze, Matthijs, Harchaoui, Zaid, & Schmid, Cordelia (2014). Category-specific video summarization. In *Proceedings of the European conference on computer vision* (pp. 540–555).

Rochan, Mrigank, Ye, Linwei, & Wang, Yang (2018). Video summarization using fully convolutional sequence networks. In *Proceedings of the European conference on computer vision* (pp. 347–363).

Song, Yale, Vallmitjana, Jordi, Stent, Amanda, & Jaimes, Alejandro (2015). Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5179–5187).

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, et al. (2017). Attention is all you need. *Vol. 30*, In *Proceedings of the advances in neural information processing systems*.

Wang, Shuai, Cong, Yang, Cao, Jun, Yang, Yunsheng, Tang, Yandong, Zhao, Huaici, et al. (2016). Scalable gastroscopic video summarization via similar-inhibition dictionary selection. *Artificial Intelligence in Medicine*, *66*, 1–13.

Wang, Zhikang, He, Lihuo, Tu, Xiaoguang, Zhao, Jian, Gao, Xinbo, Shen, Shengmei, et al. (2021). Robust video-based person re-identification by hierarchical mining. *IEEE Transactions on Circuits and Systems for Video Technology*, 1.

Wei, Huawei, Ni, Bingbing, Yan, Yichao, Yu, Huanyu, Yang, Xiaokang, & Yao, Chen (2018). Video summarization via semantic attended networks. *Vol. 32*, In *Proceedings of the AAAI conference on artificial intelligence*.

Xiao, Shuwen, Zhao, Zhou, Zhang, Zijian, Guan, Ziyu, & Cai, Deng (2020). Query-biased self-attentive network for query-focused video summarization. *IEEE Transactions on Image Processing*, *29*, 5889–5899.

Xie, Jiehang, Chen, Xuanbai, Zhang, Tianyi, Zhang, Yixuan, Lu, Shao-Ping, Cesar, Pablo, et al. (2022). Multimodal-based and aesthetic-guided narrative video summarization. *IEEE Transactions on Multimedia*, 1–15.

Xu, Minghao, Wang, Hang, Ni, Bingbing, Zhu, Riheng, Sun, Zhenbang, & Wang, Changhu (2021). Cross-category video highlight detection via set-based learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7970–7979).

Yang, Antoine, Miech, Antoine, Sivic, Josef, Laptev, Ivan, & Schmid, Cordelia (2022). Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 16442–16453).

Yeh, Ching-Feng, Wang, Yongqiang, Shi, Yangyang, Wu, Chunyang, Zhang, Frank, Chan, Julian, et al. (2021). Streaming attention-based models with augmented memory for end-to-end speech recognition. In *Proceedings of the IEEE spoken language technology workshop* (pp. 8–14).

Yuan, Yitian, Mei, Tao, Cui, Peng, & Zhu, Wenwu (2019). Video summarization by learning deep side semantic embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(1), 226–237.

Yuan, Li, Tay, Francis Eng Hock, Li, Ping, & Feng, Jiashi (2020). Unsupervised video summarization with cycle-consistent adversarial LSTM networks. *IEEE Transactions on Multimedia*, *22*(10), 2711–2722.

Zhang, Ke, Chao, Wei-Lun, Sha, Fei, & Grauman, Kristen (2016a). Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1059–1067).

Zhang, Ke, Chao, Wei-Lun, Sha, Fei, & Grauman, Kristen (2016b). Video summarization with long short-term memory. In *Proceedings of the European conference on computer vision* (pp. 766–782).

Zhang, Yunzuo, Guo, Wei, Wu, Cunyu, Li, Wei, & Tao, Ran (2023). FANet: An arbitrary direction remote sensing object detection network based on feature fusion and angle classification. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhang, Yunzuo, Kang, Weili, Liu, Yameng, & Zhu, Pengfei (2023). Joint multi-level feature network for lightweight person re-identification. In *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing* (pp. 1–5).

Zhang, Yunzuo, Song, Zhouchen, & Li, Wenbo (2023). Enhancement multi-module network for few-shot leaky cable fixture detection in railway tunnel. *Signal Processing: Image Communication*, *113*, Article 116943.

Zhang, Yunzuo, Tao, Ran, & Wang, Yue (2017). Motion-state-adaptive video summarization via spatiotemporal analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*(6), 1340–1352.

Zhang, Yunzuo, Zhang, Tian, Wu, Cunyu, & Tao, Ran (2023). Multi-scale spatiotemporal feature fusion network for video saliency prediction. *IEEE Transactions on Multimedia*.

Zhang, Shu, Zhu, Yingying, & Roy-Chowdhury, Amit K. (2016). Context-aware surveillance video summarization. *IEEE Transactions on Image Processing*, *25*(11), 5469–5478.

Zhao, Bin, Gong, Maoguo, & Li, Xuelong (2022). Hierarchical multimodal transformer to summarize videos. *Neurocomputing*, *468*, 360–369.

Zhao, Bin, Li, Xuelong, & Lu, Xiaoqiang (2017). Hierarchical recurrent neural network for video summarization. In *Proceedings of the ACM international conference on multimedia* (pp. 863–871).

Zhao, Bin, Li, Xuelong, & Lu, Xiaoqiang (2018). Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7405–7414).

Zhao, Bin, Li, Haopeng, Lu, Xiaoqiang, & Li, Xuelong (2022). Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(5), 2793–2801.

Zhao, Jiaqi, Wang, Hanzheng, Zhou, Yong, Yao, Rui, Chen, Silin, & El Saddik, Abdulmotaleb (2022). Spatial-channel enhanced transformer for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 1.

Zhong, Rui, Wang, Rui, Zou, Yang, Hong, Zhiqiang, & Hu, Min (2021). Graph attention networks adjusted Bi-LSTM for video summarization. *IEEE Signal Processing Letters*, *28*, 663–667.

Zhou, Kaiyang, Qiao, Yu, & Xiang, Tao (2018). Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. *Vol. 32*, In *Proceedings of the AAAI conference on artificial intelligence*.

Zhu, Wencheng, Lu, Jiwen, Han, Yucheng, & Zhou, Jie (2022). Learning multiscale hierarchical attention for video summarization. *Pattern Recognition*, *122*, Article 108312.

Zhu, Wencheng, Lu, Jiwen, Li, Jiahao, & Zhou, Jie (2021). DSNet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, *30*, 948–962.

Zhuang, Yueting, Rui, Yong, Huang, T. S., & Mehrotra, S. (1998). Adaptive key frame extraction using unsupervised clustering. *Vol. 1*, In *Proceedings of the international conference on image processing* (pp. 866–870).