

Contextual Correspondence Matters: Bidirectional Graph Matching for Video Summarization

Yunzuo Zhang and Yameng Liu

Shijiazhuang Tiedao University, Shijiazhuang, 050043, China
{zhangyunzuo888, liuym4647}@sina.com

Abstract. Video summarization plays a vital role in improving video browsing efficiency and has various applications in action recognition and information retrieval. In order to generate summaries that can provide key information, existing works have been proposed to simultaneously explore the contribution of both long-range and short-range temporal cues. However, they rarely consider the potential correspondence between temporal cues at different granularity within video sequences, making it insufficient to ensure detailed video understanding. In order to solve this issue, we propose a novel video summarization framework, namely Bgm4Video, based on the graph-matching mechanism, which models the potential contextualized relationship across multi-granularity temporal cues. The proposed framework is composed of two dominant components including (i) a temporal encoder (TE) that explores both coarse-grained and fine-grained contextual information within videos and (ii) a bidirectional graph transmission (BGT) module that models the interrelationship across multi-granularity temporal cues. By grasping the contextual correspondence, our method allows for further refining temporal representations to precisely pinpoint valuable segments. We demonstrate the advantage of our components through an extensive ablation study. We also evaluate our full approach on the video summarization task and demonstrate improvements over state-of-the-art on the popular benchmarks.

Keywords: Video summarization · Temporal cues · Graph matching

1 Introduction

In the age of digital media, the exponential growth of video content has ushered in an era where we are inundated with vast amounts of visual information. With the proliferation of mobile devices and the rise of platforms like YouTube, the daily influx of videos has reached staggering proportions [15]. This deluge of content has created an urgent demand for innovative solutions to help users efficiently navigate and comprehend this massive video overload [13, 16, 62].

Video summarization has emerged as a compelling answer to this challenge. By identifying and removing redundant or less informative content, video summarization allows users to swiftly grasp the most crucial information, making it

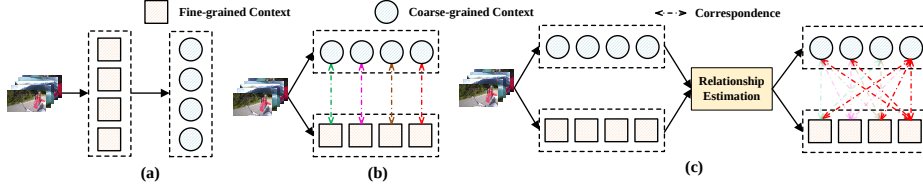


Fig. 1: Comparison between existing methods and Bgm4Video. Different from existing methods (a) and (b), our method (c) considers the contextual correspondence across multiple frames.

an invaluable tool for various applications, from content consumption to information retrieval [32], surveillance [52, 53], and beyond.

Early, traditional methods [6, 23, 28, 39] primarily rely on low-level visual or audio features, making them less effective in capturing complex contextual information and the true significance of the content [20]. Thanks to the powerful representation capability of deep learning [51], modern methods come to be developed. Long short-term memory (LSTM) [12] is usually adopted to enrich appearance features with contextual information. Although LSTM has achieved certain performance, the performance often falls short of expectations when dealing with a large number of frames [57]. To address this problem, some hierarchical structures [54, 56, 57, 59, 61] are proposed, where local contextual information is modeled before learning global temporal cues. They usually segment the entire video sequence into subsequences, which serve as local modeling units and the basis for learning long-range contextual information [8]. However, this either introduces too much noise or discards too many valuable characteristics due to inappropriate sequence segmentation [59].

On the other hand, there exist contributions [1, 17] that adopt a parallel modeling approach, where long-range and short-range contextual information is aggregated through independent pathways, using a pre-set segmentation parameter. However, they ignore the potential correspondence between temporal cues at different granularity. Simple fusion strategies make it difficult to ensure detailed video understanding. Also, due to the presence of the segmentation hyperparameter, the generalization performance is also of concern.

Inspired by the intuitive choice of modeling long-range and short-range contextual information of self-attention [42] and temporal convolution [14, 36], as well as the fact of the excellent performance of the graph to explore sophisticated structural information [34, 50], we develop a novel framework termed Bgm4Video to address video summarization. As depicted in Fig. 1, our method allows for learning wide contextual correspondence between and within coarse-grained and fine-grained temporal cues after robustly modeling the global and local contextual information of the video sequence, thereby facilitating the aggregation of more powerful contextualized representations.

More specifically, as depicted in Fig. 2, our architecture consists of two key components including a temporal encoder (TE) and a bidirectional graph trans-

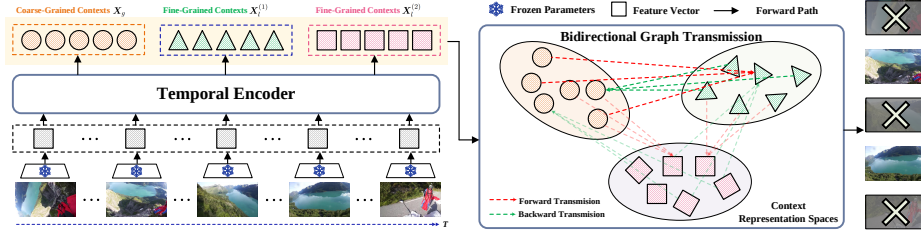


Fig. 2: Overview of the proposed method. The model first takes a video as input and extracts visual features using a pre-trained visual backbone. Our temporal encoder aims to aggregate coarse-grained contextual information $\mathbf{X}_g$ and fine-grained contextual information $\mathbf{X}_t^{(i)}$ ($i = 1, 2$). BGT performs context refinement in the context representation spaces by estimating the correspondence across multi-granularity contextual information. Lastly, our method predicts important scores given the output of BGT and generate video summaries.

mission (BGT) module. Our TE simultaneously aggregates global and local contextual information with the combination of self-attention and temporal convolution scheme, allowing for capturing rich and finer temporal cues using multiple convolution operators for accurate video understanding. BGT adopts the graph modeling scheme, which builds communication paths across graph nodes with similar contextual semantics and performs message passing.

Our contributions are summarized as follows:

(1) We reflect on the potential contextual relationships across temporal cues at different granularity, proposing a novel architecture namely Bgm4Video, which addresses the refinement problem of contextual information by resorting to a graph-matching approach. (2) We propose a bidirectional graph transmission mechanism, which enables comprehensive as well as effective message passing across multi-granularity temporal cues for grasping long-range content understanding. (3) We conduct comprehensive experiments, showing the effectiveness of our framework by achieving performance on par or better than state-of-the-art on standard datasets, showing significant improvements over previously published methods on the SumMe [9] and TVSum [39] datasets.

2 Related Work

2.1 Video Summarization

Creating excellent video summaries has remained a continual challenge, prompting researchers to investigate numerous promising methods over the years. The existing approaches can be broadly categorized into two groups: traditional methods [6, 9, 22, 31, 39, 47], which rely on conventional techniques, and deep learning-based methods [16, 18, 19, 41, 48, 49, 56, 59, 60] which harness the power of neural network models.

Traditional techniques typically employ unsupervised learning paradigms to identify and curate key frames or shots that encapsulate the core content. Cluster-based approaches [5, 6] represent an initial and straightforward approach to video summarization. In this approach, the aim 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 falls within the domain of unsupervised techniques, involving the selection of representative elements from the video to construct a summary dictionary capable of accurately reconstructing the content [29]. Nevertheless, the performance of these traditional methods remains suboptimal due to their limited representation capability, underscoring the necessity for deep learning methods.

Thanks to the success of deep learning, researchers have shifted their focus towards modern video summarization frameworks, which can be further categorized into three types: unsupervised methods, weakly supervised methods, and supervised methods. Unsupervised methods revolve around formulating criteria to measure representativeness and diversity [26, 49, 50, 56, 60]. Weakly supervised methods typically summarize videos by utilizing potentially valuable information related to the videos [3]. In contrast, supervised methods leverage manual annotations to aid in learning effective representations and often achieve superior performance [62]. To attain a comprehensive understanding of videos, many works take into consideration both global and local contextual information. Some hierarchical structures, such as HSA-RNN [59], LHMA [61], and RSGN [56], have been developed. They first model short-range temporal cues with shots and then learn long-range temporal cues within videos. However, these methods are susceptible to the negative impact of imprecise sequence segmentation [35]. While some methods [17] adopted parallel modes of learning features globally and locally, respectively, they struggle with generalization due to predefined segmentation hyperparameters. This paper presents a supervised summarization method without sequence segmentation with contingency, demonstrating effective generalization performance when processing various types of videos.

2.2 Graph Matching

Graph matching is a fundamental problem in graph pattern recognition, with the objective of identifying node correspondences between different graphs [2]. With the capability of Graph Neural Networks (GNNs) [37] to encode structural information in graphs, there has been a recent emergence of approaches that leverage GNNs to tackle the graph-matching problem. For instance, Li et al. [24] introduced a GNN-based graph-matching network to represent graphs as feature vectors, simplifying the measurement of their similarity. Li et al. [21] proposed a framework to address domain adaptive object detection, where the domain gap is reduced using graph matching. Wang et al. [43] formulated the image-text retrieval task as cross-modal scene graph matching. Cheng et al. [4] proposed a cross-modal graph matching network to achieve image-text retrieval. Soldan et al. [38] also presented a graph-matching framework to address the video grounding task, where the connectivity between video snippets and tokens is learned

using graph representation. Inspired by these works, we draw upon the fundamental concept of graph matching to learn potential semantic correspondences among contextual information within videos. Our approach specifically realizes a bidirectional graph transmission mechanism.

3 Methodology

3.1 Preliminary

Problem Definition and Overview. Given a video, our objective is to output a subsequence set, each composed of several consecutive video frames. This is challenging as it requires accurate content understanding by modeling rich contextual information within videos where these videos might have thousands of frames. The learning of powerful contextualized representations is the major focus. To address this issue, we propose a novel summarization framework based on the graph-matching mechanism, which allows for learning powerful temporal cues by modeling the potential contextual relationship across multi-granularity temporal cues. Fig. 2 shows the overview of the proposed summarization framework. In particular, our temporal encoder (TE) models both coarse-grained and fine-grained contextual information with the combination of self-attention and temporal convolution scheme. Our bidirectional graph transmission (BGT) module adopts the graph modeling approach to build communication paths and perform message passing.

Standard Attention Mechanism. This part briefly reviews the standard attention mechanism. Mathematically, the attention matrix $\boldsymbol{\vartheta} \in \mathbb{R}^{N_q \times N_c}$ is defined as $\boldsymbol{\vartheta} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)$; $\mathbf{Q} = \mathbf{Q}_t\mathbf{W}_q$, $\mathbf{K} = \mathbf{K}_t\mathbf{W}_k$, where $\mathbf{Q} \in \mathbb{R}^{N_q \times D}$ and $\mathbf{K} \in \mathbb{R}^{N_c \times D}$ are linearly projected from the query $\mathbf{Q}_t \in \mathbb{R}^{N_q \times D}$ and key $\mathbf{K}_t \in \mathbb{R}^{N_c \times D}$. The refined features $\tilde{\mathbf{V}} \in \mathbb{R}^{N_q \times D}$ can be computed via $\tilde{\mathbf{V}} = \boldsymbol{\vartheta}\mathbf{V}\mathbf{W}_o$; $\mathbf{V} = \mathbf{V}_t\mathbf{W}_v$, where the value $\mathbf{V}_t \in \mathbb{R}^{N_c \times D}$. $\mathbf{W}_{\{q,k,v,o\}} \in \mathbb{R}^{D \times D}$ indicates linear transform weights. $N_{\{q,c\}}$ denotes the length of tokens. To simplify descriptions, we define notation λ_f to get $\tilde{\mathbf{V}}$ formed by $\tilde{\mathbf{V}} = \lambda_f(\mathbf{Q}_t; \mathbf{K}_t; \mathbf{V}_t)$.

3.2 Feature Extraction

In order to ensure a relatively fair comparison with state-of-the-art methods [18, 19, 48, 54, 56, 60], we follow previous methods and use the output of the penultimate layer (pool 5) of the GoogLeNet [40] pre-trained on ImageNet [7] to represent the visual content of each frame within videos, denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T] \in \mathbb{R}^{T \times D}$, where T and D denote the duration of the input video and the dimension of visual features, respectively. These features are then used to capture contextual information described in Sec. 3.3.

3.3 Temporal Encoder

Temporal cues play a pivotal role in the comprehension of videos [25, 27, 46]. Modeling the entire video sequence offers a high-level overview of the storyline

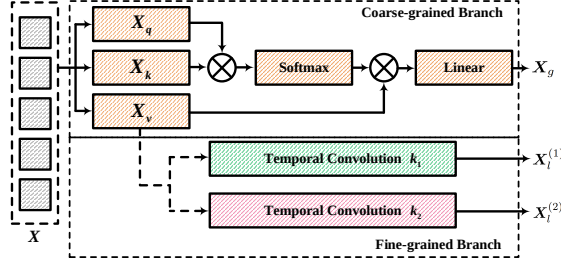


Fig. 3: Illustration of TE, which consists of a coarse-grained branch (achieved by modeling the relationship across frames) and a fine-grained branch (achieved by multiple temporal convolution operators) to aggregate long-range temporal features X_g and short-range temporal features $X_l^{(1)}$ and $X_l^{(2)}$, respectively.

while focusing on local sequences allows for an effective analysis of detailed information within specific time segments. To this end, we introduce TE, as shown in Fig. 3, comprising a coarse-grained branch and a fine-grained branch to consolidate global and local contextual information, respectively.

Coarse-grained Branch. The effectiveness of long-range modeling and computational efficiency can be assessed by considering the maximum path length and the minimum number of sequential operations [1, 42]. In both of these aspects, the self-attention mechanism offers significant advantages over LSTM. When it comes to capturing global temporal cues $X_g = [\mathbf{x}_1^g, \mathbf{x}_2^g, \dots, \mathbf{x}_T^g] \in \mathbb{R}^{T \times D}$, we employ self-attention to obtain responses at all positions in only one step, significantly avoiding the sophisticated iterative aggregation process of recursive structures. Technologically, the coarse-grained branch starts from projecting X into representation spaces $X_q \in \mathbb{R}^{T \times D}$, $X_k \in \mathbb{R}^{T \times D}$, and $X_v \in \mathbb{R}^{T \times D}$, followed by pairwise attention computation and feature aggregation, as described in Sec. 3.1 when the query, key, and value matrices are identical. Simply put, we define this branch as follows:

$$X_g = \lambda_f(X; X; X) \quad (1)$$

Fine-grained Branch. When it comes to local information, our approach places a strong emphasis on modeling temporal cues within tiny windows to gain a detailed understanding of the video storyline. We specifically opt for temporal convolution as our fine-grained context aggregator, inspired by the proven track record of excellence in extracting local features within the realm of computer vision [10]. We opt to specify more than one window size to capture more valuable information from input videos. To strike a balance between summarization effectiveness and the number of parameters, we set the size of local aggregators as $\{k_1 = 3, k_2 = 5\}$, as most contributions choose, and employ depth-separable convolution [45] to reduce the computational burden. Technically, the fine-grained stream commences with the projected matrix X_v to model short-range contextual information within the same representation latent space as the coarse-grained branch. The aggregated result $X_l^{(i)} = [\mathbf{x}_1^{l(i)}, \mathbf{x}_2^{l(i)}, \dots, \mathbf{x}_T^{l(i)}] \in \mathbb{R}^{N \times D}$

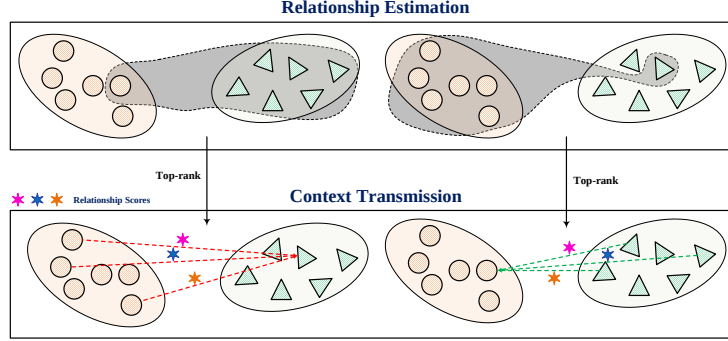


Fig. 4: Illustration of BGT. It aims to capture the potential relationship across multi-granularity contextual information within videos. BGT consists of two stages including (i) relationship estimation that is used to build passing pathways while generating correlation scores and (ii) context transmission that achieves feature aggregation by a bidirectional transmission mechanism.

according to k_i ($i = 1, 2$) can be defined as:

$$\mathbf{X}_l^{(i)} = \text{PConv}(\text{DConv}_{(k_i)}(\mathbf{X}_v)) \quad (2)$$

where $\text{PConv}(\cdot)$ and $\text{DConv}(\cdot)$ denote the pointwise and depthwise convolution.

3.4 Bidirectional Graph Transmission

Coarse-grained temporal features $\mathbf{X}_g$ and fine-grained temporal features $\mathbf{X}_l^{(i)}$ ($i = 1, 2$) learned by our temporal encoder meet and interact at the bidirectional graph transmission (BGT) layer. This layer aims to refine the obtained contextualized representations as well as grasp long-range video understanding comprehensively. It does so by modeling the potential contextual relationships across temporal cues at different granularity. Our core idea is that similar events might occur repeatedly in a video sequence. To this end, information exchange between contextual features at different positions is conducive to a comprehensive grasp of the storyline. The illustration of BGT is shown in Fig. 4, which can be divided into two stages including (i) relationship estimation and (ii) context transmission.

Relationship Estimation. Given the aggregated coarse-grained contextual information $\mathbf{X}_g$ and fine-grained contextual information $\mathbf{X}_l^{(i)}$ ($i = 1, 2$), we construct their contextual relationship leveraging the k-nearest neighbors algorithm. In order to prevent the loss of generality, we use $\mathbf{X}_g$ and $\mathbf{X}_l^{(1)}$ to explain the process in a detailed manner. Technologically, we first model bidirectional pairwise similarity relationships $\omega_i^{g \rightarrow l^{(1)}}, \omega_i^{l^{(1)} \rightarrow g} \in \mathbb{R}^T$ across i -th context and other contexts, leveraging the L_2 -norm function $\gamma(\mathbf{g}_1, \mathbf{g}_2) = \|\mathbf{g}_1 - \mathbf{g}_2\|_2$, defined as:

$$\omega_i^{g \rightarrow l^{(1)}} = \left\{ \gamma(\mathbf{x}_i^g, \mathbf{x}_j^{l^{(1)}}) | j = 1, 2, \dots, T \right\} \quad (3)$$

$$\omega_i^{l^{(1)} \rightarrow g} = \left\{ \gamma(\mathbf{x}_i^{l^{(1)}}, \mathbf{x}_j^g) | j = 1, 2, \dots, T \right\} \quad (4)$$

The most similar temporal cues are established by performing the top-rank operator on these nodes, formed by

$$\mathbf{I}_i^{g \rightarrow l^{(1)}} = \text{top-rank}(\omega_i^{g \rightarrow l^{(1)}}, L) \quad (5)$$

$$\mathbf{I}_i^{l^{(1)} \rightarrow g} = \text{top-rank}(\omega_i^{l^{(1)} \rightarrow g}, L) \quad (6)$$

where $\mathbf{I}_i^{g \rightarrow l^{(1)}}$ and $\mathbf{I}_i^{l^{(1)} \rightarrow g}$ are the index sets, which select the positions that are similar to the current position. L is the number of neighbors. L is used to control the number of neighbor nodes.

Context Transmission. Subsequently, for each direction, BGT adopts the **Softmax** function to normalize corresponding scores according to the select positions, which are used to adaptively achieve context aggregation. Then, we sum temporal features up by weighting, followed by a residual connection to perform information exchange. With bidirectional aggregation features, based on their summation, BGT adopts convolution using a split-transform-merge strategy [44] to further refine features. At this point, only the description of the information exchange between coarse-grained features $\mathbf{X}_g$ and the first fine-grained features $\mathbf{X}_l^{(1)}$ is provided, and the same operation is also applied between $\mathbf{X}_g$ and $\mathbf{X}_l^{(2)}$ as well as $\mathbf{X}_l^{(1)}$ and $\mathbf{X}_l^{(2)}$. We conduct experiments to demonstrate the effectiveness of this thinking in Sec. 4.

3.5 Summarization Generation

Given the output of BGT, we perform element-wise summation between it and features $\mathbf{X}$ to retain original visual information. Subsequently, our prediction head, formed by **Linear** $\rightarrow$ **ReLU** $\rightarrow$ **Dropout** $\rightarrow$ **LayerNorm** $\rightarrow$ **Linear** $\rightarrow$ **Sigmoid**, predicts frame-level importance scores $\mathbf{S} = [s_1, s_2, \dots, s_T]$. Following previous works [48], shot-level scores are computed by averaging frame-level importance scores within detected shots obtained by the KTS algorithm [35]. Next, according to a length percentage threshold (*e.g.*, 15% of the total length), Bgm4Video generates video summaries by defining a 0/1 knapsack problem and addressing this using dynamic programming.

3.6 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:

Table 1: Characteristics statistics of the SumMe [9], TVSum [39], OVP [6], and YouTube [6] datasets.

Dataset	Number of Videos	Number of Users	Annotation Type	Duration of Videos		
				Min	Max	Avg
SumMe [9]	25	15-18	importance score	32s	5min 24s	2min 26s
TVSum [39]	50	20	importance score	1min 23s	10min 47s	3min 55s
OVP [6]	50	5	key frame	9s	9min 32s	1min 38s
YouTube [6]	39	5	key frame	46s	3min 29s	3min 16s

$$\mathcal{L}(\theta) = \frac{1}{T} \sum_{i=1}^T (s_i - y_i)^2 \quad (7)$$

where θ is the parameters. s_i and y_i are the predicted human-created importance score of i -th frame, respectively.

4 Experiments

We conduct experiments to verify the effectiveness of our method. We first provide a detailed introduction to the experimental settings, including the dataset, implementation details, and evaluation metrics. Then, we compare the experimental results of the proposed method with existing methods under different evaluation manners. Then, we conduct ablation experiments to verify the impact of different modules. Finally, we provide some visualization results.

4.1 Experimental Settings

Datasets We evaluate our method on two benchmark datasets including SumMe [9] and TVSum [39]. 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 minutes. 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 minutes. In addition, we employ YouTube [6] and OVP [6] to augment our training data. The detailed information of used datasets is summarized in Table 1.

Implementation Details Following previous methods [19, 48, 60], videos are sub-sampled at 2fps to reduce temporal redundancy in advance, as the visual content of adjacent frames is extremely similar. The number of neighbors L is set to 3. The framework is trained by Adam for 300 epochs with batch size 1 and a learning rate of 5×10^{-5} . The same as [19], we adopt standard 5-fold cross-validation to ensure that each video participates in the testing procedure, and thus generates more reliable experimental results.

Table 2: Comparisons with other state-of-the-art methods on the SumMe and TVSum datasets under the standard evaluation setting in the F-score (%) evaluation. We also present the number of parameters (M) and the test method of different contributions.

Method	SumMe $\uparrow$	TVSum $\uparrow$	Params $\downarrow$	Evaluation
ERSUM [22]	0.431	0.594	-	-
MSDS-CC [31]	0.406	0.523	-	-
vsLSTM [48]	0.376	0.542	2.63	5 Random
dppLSTM [48]	0.386	0.547	2.63	5 Random
SUM-GAN [30]	0.417	0.563	295.86	5 Random
DR-DSN [60]	0.421	0.581	2.63	5 FCV
SUM-FCN [36]	0.475	0.568	36.58	M Random
HSA-RNN [59]	0.423	0.587	4.20	5 Random
ACGAN [11]	0.472	0.594	31.51	5 FCV
CSNet [17]	0.486	0.585	-	5 FCV
AVRN [55]	0.441	0.597	-	5 Random
RSGN [56]	0.450	0.601	-	5 Random
LMHA [61]	0.511	0.610	-	5 Random
HMT [58]	0.441	0.601	-	5 Random
VJMHT [19]	0.506	0.609	35.44	5 FCV
Bgm4Video	0.517	0.612	7.62	5 FCV

Evaluation Metrics Following state-of-the-art methods [19, 48, 57, 59–61], we first report the F-score performance on the SumMe and TVSum datasets, which is defined as:

$$\text{Precision} = \frac{\text{overlapped duration of } \mathbf{F}_p \text{ and } \mathbf{F}_h}{\text{duration of } \mathbf{F}_p} \quad (8)$$

$$\text{Recall} = \frac{\text{overlapped duration of } \mathbf{F}_p \text{ and } \mathbf{F}_h}{\text{duration of } \mathbf{F}_h} \quad (9)$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

where $\mathbf{F}_p$ and $\mathbf{F}_h$ represent the predicted and annotated summaries, respectively.

Based on the limitation of the F-score evaluation [33], we follow previous methods [19, 56, 58] and report correlation coefficients including Kendall’s τ and Spearman’s ρ for further evaluating our method. They are used to measure the agreement between predicted scores and annotated scores by humans and make the superiority of the proposed method more reliable.

4.2 Comparisons with State-of-the-Arts

In this section, we compare our approach to state-of-the-art methods in the video summarization task. The comparison results in the F-score evaluation are first reported, which is followed by the results in the rank-based evaluation.

Table 3: Comparisons with other state-of-the-art methods on the SumMe and TVSum datasets under canonical (Can.), augmented (Aug.), and transfer (Trans.) settings in the F-score (%) evaluation.

Methods	SumMe $\uparrow$			TVSum $\uparrow$		
	Can.	Aug.	Tran.	Can.	Aug.	Tran.
vsLSTM [48]	0.376	0.416	0.407	0.542	0.579	0.569
dppLSTM [48]	0.386	0.429	0.418	0.547	0.596	0.587
SUM-GAN [30]	0.417	0.436	-	0.563	0.612	-
DR-DSN [60]	0.421	0.439	0.426	0.581	0.598	0.589
SUM-FCN [36]	0.475	0.511	0.441	0.568	0.592	0.582
HSA-RNN [59]	0.423	0.421	-	0.587	0.598	-
CSNet [17]	0.486	0.487	0.441	0.585	0.571	0.574
AVRN [55]	0.441	0.449	0.432	0.597	0.605	0.587
RSGN [56]	0.450	0.457	0.440	0.601	0.611	0.600
LMHA [61]	0.511	0.521	0.454	0.610	0.615	0.551
HMT [58]	0.441	0.448	-	0.601	0.603	-
VJMHT [19]	0.506	0.517	0.464	0.609	0.619	0.589
Bgm4Video	0.517	0.541	0.469	0.612	0.617	0.574

Comparisons in F-Score Evaluation Comparison Under the Standard Setting. We compare our method with previous works in the widely-used F-score evaluation under the standard evaluation setting and the results are shown in Table 2.

From the reported values, we can observe that our Bgm4Video achieves an extremely comparable summarization performance. Compared with previous hierarchical frameworks such as HSA-RNN [57], HTM [58], RSGN [56], LMHA [61], our method adopts parallel modeling approach to capture both global and local temporal cues, effectively avoiding the negative impact caused by inaccurate sequence segmentation. Compared with CSNet [17] that adopts the parallel modeling manner, our method outperforms them, which might be attributed to the better generalization performance due to not requiring manual specification of the number of segments in our framework. Moreover, the parameters of some methods are also displayed and Bgm4Video achieves a good trade-off between the F-score and parameters.

Comparison Under Other Evaluation Settings. We further conduct experiments under the augmented and transfer settings on the SumMe and TVSum datasets along with additional OVP and YouTube datasets, as illustrated in Table 3. We can observe that Bgm4Video achieves an encouraging performance. Specifically, our method achieves state-of-the-art performance in some of these settings, while the results obtained in other settings are not satisfactory. We believe that there may be overfitting during the training process, which also indicates the necessity of enhancing the scalability and transferability of the model.

Table 4: Comparisons with other state-of-the-art methods in the rank-based evaluation. This experiment uses the TVSum [39] dataset under the standard setting.

Method	Kendall’s $\tau \uparrow$	Spearman’s $\rho \uparrow$
Random [33]	0.000	0.000
Human [33]	0.177	0.204
dppLSTM [48]	0.042	0.055
DR-DSN [60]	0.020	0.026
HSA-RNN [59]	0.082	0.088
AVRN [55]	0.096	0.107
RSGN [56]	0.083	0.090
HMT [58]	0.096	0.107
VJMHT [19]	0.097	0.105
Bgm4Video	0.169	0.221

Table 5: Ablation study on context granularity.

Exp.	Granularity Settings		SumMe	TVSum
	Coarse <i>only</i>	Fine <i>only</i>		
1	✓		0.487	0.597
2		✓	0.511	0.605
3	✓	✓	0.517	0.612

Comparisons in Rank-based Evaluation We compare our methods with state-of-the-art methods using the rank correlation coefficients, Spearman’s ρ and Kendall’s τ . The results are shown in Table 4.

As can be seen, our method outperforms existing methods to a large extent, which means the proposed method can model the relative importance among frames more accurately than previous works. The experimental results in this more reliable evaluation metric provide strong evidence of the effectiveness of our method.

4.3 Ablation Study

Our framework consists of two dominant components including a temporal encoder and a bidirectional graph transmission module. To demonstrate the effectiveness of each part, we conduct an ablation study, which can be seen in the following.

Study on Granularity. As mentioned in prior parts, coarse-grained and fine-grained contextual information plays a crucial role in video understanding. To this end, we check their contributions in summarizing videos, as shown in Table 5. Coarse (Coarse-grained Branch) *only* and Fine (Fine-grained Branch) *only*

Table 6: Ablation study on transmission direction.

Exp.	Direction Settings		SumMe	TVSum
	Forward <i>only</i>	Backward <i>only</i>		
1	✓		0.502	0.608
2		✓	0.507	0.608
3	✓	✓	0.517	0.612

Table 7: Ablation study on transmission position.

Position Settings			SumMe	TVSum
$X_g \leftrightarrow X_l^{(1)}$	$X_g \leftrightarrow X_l^{(2)}$	$X_l^{(1)} \leftrightarrow X_l^{(2)}$		
			0.508	0.604
✓			0.510	0.607
✓	✓		0.514	0.610
✓	✓	✓	0.517	0.612

indicate that the framework only contains coarse-grained and fine-grained contextual information. As can be seen, the summarization performance achieves the best while coarse-grained and fine-grained temporal information are simultaneously enriched into visual features (Exp. 3). This is because they allow for providing more valuable information for video understanding.

Study on Transmission Direction. To verify the effectiveness of the bidirectional transmission mechanism in BGT, we conduct ablation experiments to effect of unidirectional transmission. As Table 6 displays, our framework (Exp. 3) using the bidirectional transmission (simultaneously including forward and backward transmission) achieves the best video summarization performance, validating the necessity of comprehensive message passing from two directions.

Study on Transmission Position. We explore the impact of BGT when applied in different positions to verify the positive role of our graph-matching mechanism in information transmission. The experimental results are shown in Table 7, where (for example) $X_g \leftrightarrow X_l^{(1)}$ denotes BGT is adopted between X_g and $X_l^{(1)}$, otherwise use summation for information exchange. We can conclude that our full architecture clearly outperforms other situations, which proves the advantage of BGT in video understanding by information transmission across multi-granularity contextual information.

4.4 Visualization

To visually demonstrate the effectiveness of our approach, we present qualitative results of Bgm4Video in Fig. 5. These examples show that our method excels at identifying the most significant shots, as indicated by their high-importance

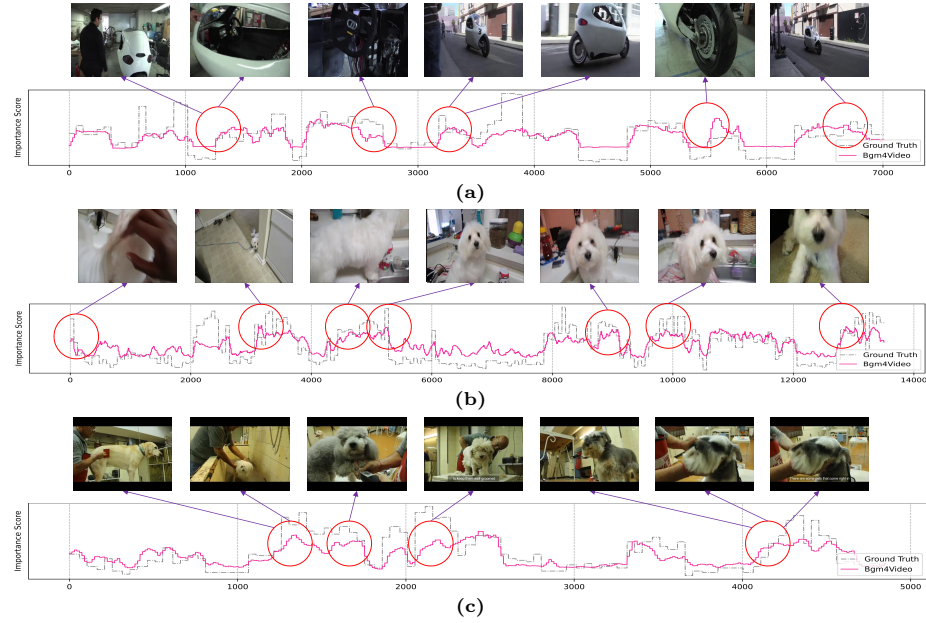


Fig. 5: Qualitative results. We show some qualitative examples on the TVSum [39] dataset. The predicted importance scores of our framework and the ground truth importance scores are shown by pink and gray lines, respectively.

scores. The results clearly highlight the effectiveness of our method in modeling the relative importance of videos.

5 Conclusion

In this paper, we take the first attempt to reflect on the potential interrelationship across multi-granularity contextual information and address this aspect by resorting to a graph-matching approach. Based on our Bgm4Video summarization framework, both coarse-grained and fine-grained contextual information is simultaneously captured by the temporal encoder, a product that combines self-attention and multiple temporal convolutions, in a parallel modeling approach. Our bidirectional graph transmission mechanism performs sophisticated information exchange across these coarse-grained and fine-grained contexts by considering their contextual correspondence. Extensive experiments have verified the effectiveness of our contributions.

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 (No. F2022210007, No. F2017210161), the Science and Technology Project of Hebei Education Department (No. ZD2022100, No. QN2017132), the Central Guidance on Local Science and Technology Development Fund (No. 226Z0501G).

References

1. Apostolidis, E., Balaouras, G., Mezaris, V., Patras, I.: Combining global and local attention with positional encoding for video summarization. In: 2021 IEEE International Symposium on Multimedia. pp. 226–234 (2021)
2. Caetano, T.S., McAuley, J.J., Cheng, L., Le, Q.V., Smola, A.J.: Learning graph matching. *IEEE transactions on pattern analysis and machine intelligence* **31**(6), 1048–1058 (2009)
3. Cai, S., Zuo, W., Davis, L.S., Zhang, L.: Weakly-supervised video summarization using variational encoder-decoder and web prior. In: Proceedings of the European Conference on Computer Vision. pp. 184–200 (2018)
4. Cheng, Y., Zhu, X., Qian, J., Wen, F., Liu, P.: Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* **18**(4), 1–23 (2022)
5. Chu, W.S., Song, Y., Jaimes, A.: Video co-summarization: Video summarization by visual co-occurrence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3584–3592 (2015)
6. De Avila, S.E.F., Lopes, A.P.B., da Luz Jr, A., de Albuquerque Araújo, A.: Vsum: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters* **32**(1), 56–68 (2011)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
8. Fu, H., Wang, H.: Self-attention binary neural tree for video summarization. *Pattern Recognition Letters* **143**, 19–26 (2021)
9. Gygli, M., Grabner, H., Riemenschneider, H., Gool, L.V.: Creating summaries from user videos. In: Proceedings of the European Conference on Computer Vision. pp. 505–520 (2014)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. He, X., Hua, Y., Song, T., Zhang, Z., Xue, Z., Ma, R., Robertson, N., Guan, H.: Unsupervised video summarization with attentive conditional generative adversarial networks. In: Proceedings of the ACM International Conference on Multimedia. pp. 2296–2304 (2019)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Computation* **9**(8), 1735–1780 (1997)
13. Huang, C., Wang, H.: A novel key-frames selection framework for comprehensive video summarization. *IEEE Transactions on Circuits and Systems for Video Technology* **30**(2), 577–589 (2020)
14. Huang, S., Li, X., Zhang, Z., Wu, F., Han, J.: User-ranking video summarization with multi-stage spatio-temporal representation. *IEEE Transactions on Image Processing* **28**(6), 2654–2664 (2018)
15. Hussain, T., Muhammad, K., Ding, W., Lloret, J., Baik, S.W., de Albuquerque, V.H.C.: A comprehensive survey of multi-view video summarization. *Pattern Recognition* **109**, 107567 (2021)
16. Ji, Z., Xiong, K., Pang, Y., Li, X.: Video summarization with attention-based encoder-decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology* **30**(6), 1709–1717 (2020)

17. Jung, Y., Cho, D., Kim, D., Woo, S., Kweon, I.S.: Discriminative feature learning for unsupervised video summarization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8537–8544 (2019)
18. Li, H., Ke, Q., Gong, M., Drummond, T.: Progressive video summarization via multimodal self-supervised learning. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 5584–5593 (2023)
19. Li, H., Ke, Q., Gong, M., Zhang, R.: Video joint modelling based on hierarchical transformer for co-summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(3), 3904–3917 (2023)
20. Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., Shao, L.: Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition* **111**, 107677 (2021)
21. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5291–5300 (2022)
22. Li, X., Zhao, B., Lu, X.: A general framework for edited video and raw video summarization. *IEEE Transactions on Image Processing* **26**(8), 3652–3664 (2017)
23. Li, Y., Merialdo, B.: Multi-video summarization based on video-mm. In: 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10. pp. 1–4 (2010)
24. Li, Y., Gu, C., Dullien, T., Vinyals, O., Kohli, P.: Graph matching networks for learning the similarity of graph structured objects. In: International conference on machine learning. pp. 3835–3845 (2019)
25. Liu, M., Jin, S., Yao, C., Lin, C., Zhao, Y.: Temporal consistency learning of interframes for video super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology* **33**(4), 1507–1520 (2022)
26. Liu, T., Meng, Q., Huang, J.J., Vlontzos, A., Rueckert, D., Kainz, B.: Video summarization through reinforcement learning with a 3d spatio-temporal u-net. *IEEE Transactions on Image Processing* **31**, 1573–1586 (2022)
27. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3042–3051 (2022)
28. Mademlis, I., Tefas, A., Pitas, I.: A salient dictionary learning framework for activity video summarization via key-frame extraction. *Information Sciences* **432**, 319–331 (2018)
29. Mademlis, I., Tefas, A., Pitas, I.: A salient dictionary learning framework for activity video summarization via key-frame extraction. *Information Sciences* **432**, 319–331 (2018)
30. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 202–211 (2017)
31. Meng, J., Wang, S., Wang, H., Yuan, J., Tan, Y.P.: Video summarization via multi-view representative selection. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 1189–1198 (2017)
32. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23023–23033 (2023)

33. Otani, M., Nakashima, Y., Rahtu, E., Heikkila, J.: Rethinking the evaluation of video summaries. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7596–7604 (2019)
34. Park, J., Lee, J., Kim, I.J., Sohn, K.: Sumgraph: Video summarization via recursive graph modeling. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. pp. 647–663 (2020)
35. Potapov, D., Douze, M., Harchaoui, Z., Schmid, C.: Category-specific video summarization. In: *Proceedings of the European Conference on Computer Vision*. pp. 540–555 (2014)
36. Rochan, M., Ye, L., Wang, Y.: Video summarization using fully convolutional sequence networks. In: *Proceedings of the European Conference on Computer Vision*. pp. 347–363 (2018)
37. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. *IEEE transactions on neural networks* **20**(1), 61–80 (2008)
38. Soldan, M., Xu, M., Qu, S., Tegner, J., Ghanem, B.: Vlg-net: Video-language graph matching network for video grounding. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 3224–3234 (2021)
39. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: Tvsum: Summarizing web videos using titles. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5179–5187 (2015)
40. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015)
41. Tu, Z., Xie, W., Dauwels, J., Li, B., Yuan, J.: Semantic cues enhanced multi-modality multistream cnn for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology* **29**(5), 1423–1437 (2019)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems*. vol. 30 (2017)
43. Wang, S., Wang, R., Yao, Z., Shan, S., Chen, X.: Cross-modal scene graph matching for relationship-aware image-text retrieval. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 1508–1517 (2020)
44. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. *ACM Transactions On Graphics* **38**(5), 1–12 (2019)
45. Wu, F., Fan, A., Baevski, A., Dauphin, Y.N., Auli, M.: Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430* (2019)
46. Xu, M., Wang, H., Ni, B., Zhu, R., Sun, Z., Wang, C.: Cross-category video highlight detection via set-based learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 7970–7979 (2021)
47. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Summary transfer: Exemplar-based subset selection for video summarization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1059–1067 (2016)
48. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: *Proceedings of the European Conference on Computer Vision*. pp. 766–782 (2016)
49. Zhang, Y., Liu, Y., Kang, W., Zheng, Y.: Mar-net: Motion-assisted reconstruction network for unsupervised video summarization. *IEEE Signal Processing Letters* (2023)

50. Zhang, Y., Liu, Y., Zhu, P., Kang, W.: Joint reinforcement and contrastive learning for unsupervised video summarization. *IEEE Signal Processing Letters* **29**, 2587–2591 (2022)
51. Zhang, Y., Song, Z., Li, W.: Enhancement multi-module network for few-shot leaky cable fixture detection in railway tunnel. *Signal Processing: Image Communication* p. 116943 (2023)
52. Zhang, Y., Tao, R., Wang, Y.: Motion-state-adaptive video summarization via spatiotemporal analysis. *IEEE Transactions on Circuits and Systems for Video Technology* **27**(6), 1340–1352 (2017)
53. Zhang, Y., Zhang, J., Liu, R., Zhu, P., Liu, Y.: Key frame extraction based on quaternion fourier transform with multiple features fusion. *Expert Systems with Applications* **216**, 119467 (2023)
54. Zhao, B., Gong, M., Li, X.: Audiovisual video summarization. *IEEE Transactions on Neural Networks and Learning Systems* pp. 1–8 (2021)
55. Zhao, B., Gong, M., Li, X.: Audiovisual video summarization. *IEEE Transactions on Neural Networks and Learning Systems* **34**(8), 5181–5188 (2023)
56. Zhao, B., Li, H., Lu, X., Li, X.: Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(5), 2793–2801 (2022)
57. Zhao, B., Li, X., Lu, X.: Hierarchical recurrent neural network for video summarization. In: *Proceedings of the ACM International Conference on Multimedia*. pp. 863–871 (2017)
58. Zhao, Bin and Gong, Maoguo and Li, Xuelong: Hierarchical multimodal transformer to summarize videos. *Neurocomputing* **468**, 360–369 (2022)
59. Zhao, Bin and Li, Xuelong and Lu, Xiaoqiang: Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7405–7414 (2018)
60. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 32 (2018)
61. Zhu, W., Lu, J., Han, Y., Zhou, J.: Learning multiscale hierarchical attention for video summarization. *Pattern Recognition* **122**, 108312 (2022)
62. Zhu, W., Lu, J., Li, J., Zhou, J.: Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing* **30**, 948–962 (2020)