Video summarization via global feature difference optimization*

ZHANG Yunzuo and LIU Yameng**

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

(Received 15 December 2022; Revised 13 February 2023)

Video summarization aims at selecting valuable clips for browsing videos with high efficiency. Previous approaches typically focus on aggregating temporal features while ignoring the potential role of visual representations in summarizing videos. In this paper, we present a global difference-aware network (GDANet) that exploits the feature difference across frame and video as guidance to enhance visual features. Initially, a difference optimization module (DOM) is devised to enhance the discriminability of visual features, bringing gains in accurately aggregating temporal cues. Subsequently, a dual-scale attention module (DSAM) is introduced to capture informative contextual information. Eventually, we design an adaptive feature fusion module (AFFM) to make the network adaptively learn context representations and perform feature fusion effectively. We have conducted experiments on benchmark datasets, and the empirical results demonstrate the effectiveness of the proposed framework.

Document code: A **Article ID:** 1673-1905(2023)09-0570-7

DOI https://doi.org/10.1007/s11801-023-2212-0

With the growing prevalence of smart mobile devices, it is convenient to acquire videos in our daily life. Massive videos can be easily uploaded to the network, which inevitably leads to the trouble that obtaining meaningful information in lengthy videos is laborious and tedious since humans have to watch the entire video content earnestly[1]. Consequently, how to effectively analyze and store these lengthy video data has become a research hotspot recently[2].

As an effective method to help viewers quickly browse and understand video content, video summarization has received extensive attention in computer vision. Its purpose is to remove redundant parts and generate a concise synopsis that can provide comprehensive and valuable information[3,4]. To date, video summarization can be realized by selecting a set of frames or shots. Compared to the former, shot-based video summarization methods is capable of conveying dramatic information and thus become a major topic in summarizing videos. Therefore, selecting key shots from an input video is precisely what this paper studies.

Existing approaches can be roughly categorized into traditional approaches and modern approaches. Traditional approaches usually generate summaries by using regular tips, such as clustering algorithm and dictionary learning. For example, CHU et al[5] considered the video topic and accordingly proposed a topic-related joint clustering algorithm to generate video summarization. MEI et al[6] put forward a video summarization method based on $L_{2,0}$ sparse dictionary selection. However, these approaches rarely take the temporal cues into account, confronting a performance bottleneck in summarizing videos.

Benefiting from the widespread success of deep learning, many modern approaches have been proposed and achieved significant progress. Generally, modern approaches take a frame sequence as input and output a set of scores indicating one frame's importance. Consequently, video summarization is usually perceived as a sequence-to-sequence learning task. In order to process variable-range frames, ZHANG et al[7] adopted bi-directional long short-term memory network to capture the contextual information, additionally combining a determinantal point process (DPP) module to derive representative video summaries. Nevertheless, recursive neural networks (RNNs) may perform poorly once the video sequence exceeds 80 frames[8]. ZHAO et al[9] designed a hierarchical recurrent neural network by aggregating contextual information within and across shots. On its basis, ZHAO et al[10] introduced a sliding window based shot boundary detection method to avoid the limitation caused by fixed length segmentation. JUNG et al[11] proposed a two-stream network to jointly consider a local and a global view of input features. FU et al[12] devised a self-attention binary neural tree to evaluate shots from different aspects. However, these methods focus on

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

** E-mail: zhangyunzuo888@sina.com

boosting the capability of mining contextual information while ignoring the effect caused by input features.

Aiming at enhancing the representation of visual information, KANAFANI et al[13] made an attempt to summarize videos by providing rich features extracted from two pre-trained networks. Despite a performance improvement, it complicated the process of extracting visual features. In this paper, we present a global difference-aware network (GDANet), which tries to enhance the discriminability of visual features by exploring the feature difference between frame and video content. Different from Ref.[13], this paper goes deeper into promoting informative feature learning, instead of providing multi-source visual features, thus boosting the conciseness of the proposed framework. The proposed framework is capable of extracting rich and representative contextual features, which greatly improves the summarization performance.

The overall structure is depicted in Fig.1. Given an input video, a convolution neural network (CNN) first extracts visual features for all frames. These features are subsequently fed into the proposed difference optimization module (DOM) for feature enhancement, which is followed by the dual-scale attention module (DSAM) and adaptive feature fusion module (AFFM) for mining contextual information and context fusion, respectively. Finally, the importance score predictor is utilized to estimate the importance. The orange bounding boxes are the selected summary. Specifically, given a video containing N frames in total, we employ GoogLeNet[14] to extract visual features denoted as $\boldsymbol{F}=\{\boldsymbol{f}_i \in \mathbb{R}^d\}_{i=1}^N$, where $\boldsymbol{f}_i$ reflects the semantic information of the ith frame, and d=1 024 is the dimension.

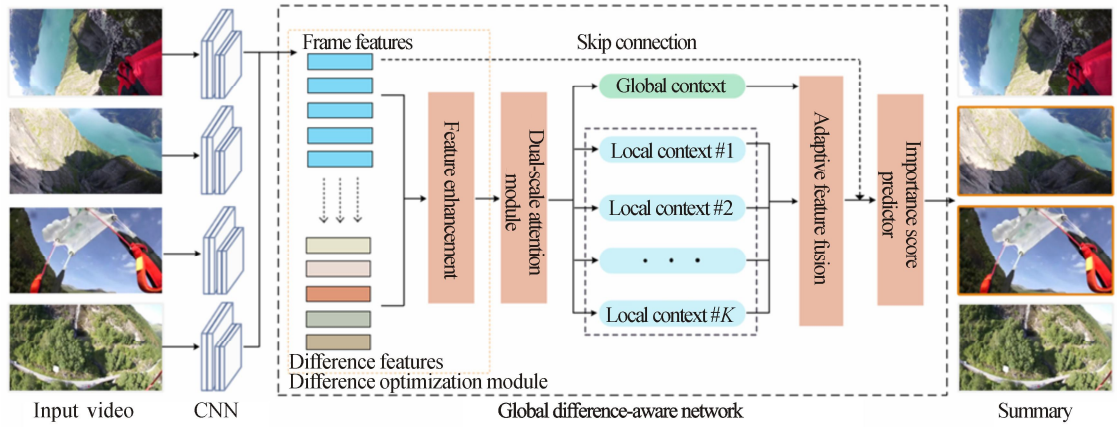


Fig.1 Overview of GDANet

To obtain discriminative feature representations, DOM is devised, which aims to enhance input features by making each vector aware of its semantic difference across the frame and video content. Mathematically, according to these feature vectors, we calculate difference features $\boldsymbol{D}=\{\boldsymbol{d}_i \in \mathbb{R}^d\}_{i=1}^N$ with respect to video content representation. The calculation process can be formulated as

$$\boldsymbol{d}_i = \left|\boldsymbol{f}_i - \boldsymbol{\varepsilon}\right|, \tag{1}$$

where $\boldsymbol{d}_i$ essentially denotes the semantic difference between the ith position and the overall video content. $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ denotes the video-level feature vector adopted to represent the video content, which particularly can be defined as the average of all frame-level visual features. The formula is represented as

$$\boldsymbol{\varepsilon} = \sum_{i=1}^{N} \boldsymbol{f}_i / N. \tag{2}$$

We perform linear transformation for feature difference $\boldsymbol{D}$ and visual feature $\boldsymbol{F}$ respectively to enhance the expression capability of our network and subsequently conduct element-wise addition to achieve intermediate feature embedding. The features are successively fed into fully connected layer and a Softmax function to generate difference attention $\boldsymbol{V}=\{v_i\}_{i=1}^N$. Besides, an element-wise addition is also performed, which can be explained by the fact that it is not only effective to promote discriminability, but also to endow features with latent position information. The optimized features $\boldsymbol{R}=\{\boldsymbol{r}_i \in \mathbb{R}^d\}_{i=1}^N$ can be formulated as

$$\boldsymbol{r}_i = \left(v_i \otimes \boldsymbol{f}_i\right) \oplus \boldsymbol{d}_i, \tag{3}$$

where $\otimes$ and $\oplus$ stand for element-wise production and addition operator, respectively. Leveraging the DOM, our model is capable of learning more effective feature representations for accurately capturing contextual information.

Subsequently, we propose DSAM, which consists of a global pathway and a local pathway for the purpose of mining multiscale temporal features. It is grounded on multi-head self-attention[15], which is capable of avoiding the problem of historical information decay in RNNs. The illustration of multi-head self-attention and attention module is shown in Fig.2.

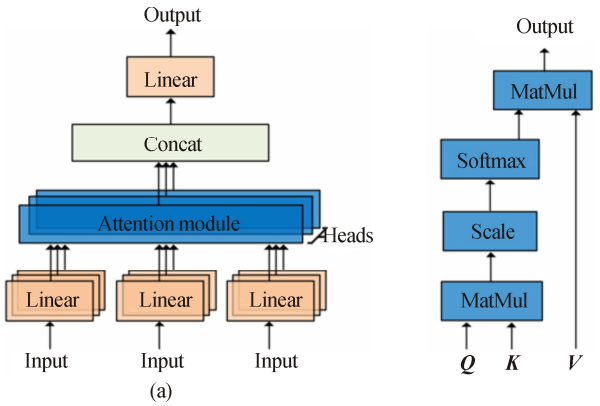


Fig.2 Illustration of (a) multi-head self-attention and (b) attention module

Fig.3 presents the core idea of DSAM. In terms of the global pathway, we conduct self-attention on the entire video sequence and obtain the coarse-grained contextual information $\boldsymbol{G} \in \mathbb{R}^{N\times d}$, which reveals the long-distance temporal cues of input videos. It is crucial to understand video content since human is likely to get a summary after an overview of a video. About the local pathway, the entire video sequence is uniformly divided into K non-overlapping subsequences denoted as $\{\boldsymbol{R}_i \in \mathbb{R}^{(N/K)\times d}\}_{i=1}^{K}$. The multi-head self-attention is separately performed on each subsequence for the purpose of mining short-distance contextual information $\boldsymbol{L} \in \mathbb{R}^{N\times d}$, which provides fine-grained contextual information. The formula can be represented as

$$\boldsymbol{G} = \xi(\boldsymbol{R}), \tag{4}$$

$$\boldsymbol{L} = \sum_{i=1}^{K} \xi(\boldsymbol{R}_i), \tag{5}$$

where $\xi(\cdot)$ is the multi-head self-attention function. In this paper, we set K as 3 by conducting parameter analysis to achieve the best summarization performance.

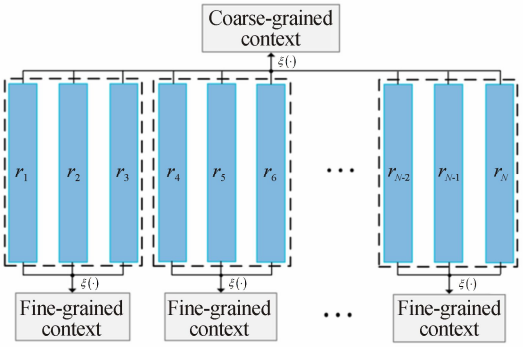


Fig.3 Illustration of DSAM

After that, we conduct feature fusion by our proposed AFFM, the purpose of which is to adaptively learn representations and achieve effective and subsequent feature fusion. This is motivated by the fact that there exists a significant semantic difference between global and local contextual information. The pipeline is shown in Fig.4. Mathematically, the low-level fusion feature is computed by element-wise addition operation between global and local contextual information. Then we compute the deep features $\boldsymbol{P}^d = \{\boldsymbol{p}_i^d \in \mathbb{R}^d\}_{i=1}^{N}$ by two fully connected layers, and the global features $\boldsymbol{P}^g \in \mathbb{R}^d$ are computed by global average pooling. The adaptive fusion coefficient matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N\times d}$ is obtained by an element-wise addition and a sigmoid layer, guiding our model to learn effective context features by element-wise production along the channel dimension. This can be written as

$$\boldsymbol{H} = \boldsymbol{G} \otimes \boldsymbol{\Phi} \oplus \boldsymbol{L} \otimes (\boldsymbol{1} - \boldsymbol{\Phi}), \tag{6}$$

where $\boldsymbol{H} \in \mathbb{R}^{N\times d}$ is the final outcome, which essentially embodies global and local contextual information and can provide valuable information for importance prediction.

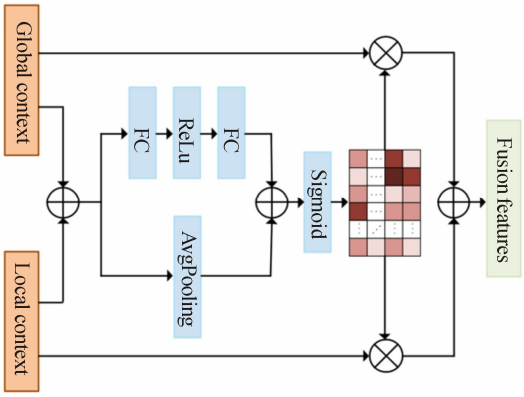


Fig.4 Illustration of AFFM

In addition to the above module, our approach also includes an importance score predictor to predict importance scores for each frame. The structure is shown in Fig.5. The final fully connected layer aims to map the features with a dimension of d to the importance score set $\boldsymbol{S} = \{s_i\}_{i=1}^{N}$, where s_i reflects the importance of the ith frame.

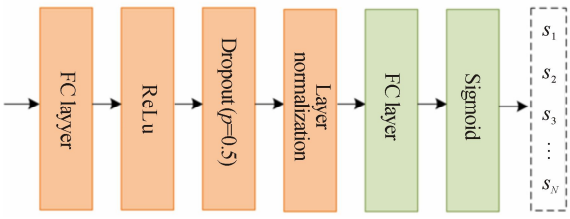


Fig.5 Illustration of importance score predictor

During the training process, we adopt mean square error (*MSE*) as the loss function, aiming at reducing the distribution inconsistency between generated summaries and manual annotations. Mathematically, the loss $\mathcal{L}(\boldsymbol{\theta})$ can be formulated as

$$\mathcal{L}(\boldsymbol{\theta}) = \|\boldsymbol{O} - \boldsymbol{S}\|_2^2, \tag{7}$$

where $\boldsymbol{\theta}$ is the parameters in our model and $\boldsymbol{O}$ stands for ground truth. To select the most valuable key shots of input videos, we employ kernel temporal segmentation[16] to detect shot boundary points and convert frame-level importance scores into shot-level importance scores by

taking an average of scores within a shot. Then, a knapsack problem is created, which can be mathematically defined as

$$\max \sum_{i=1}^{N} a_i u_i \text{ , s.t. } \sum_{i=1}^{N} a_i l_i \le 0.15 \times L \text{ ,} \tag{8}$$

where $a_i \in \{0,1\}$ stands for whether the ith shot is selected into summary. u_i and l_i are the shot-level importance score and length. L denotes the total length of an input video. The summary can be obtained by rearranging the shots limited with a_i=1.

In order to demonstrate the effectiveness of our proposed summarization framework, we conduct experiments on the most popular benchmark datasets, including SumMe[17] and TVSum[18]. The SumMe dataset contains 25 videos from YouTube involving food and sports, etc., and each video is annotated by 15—18 users, and the video duration varies from 1 min to 6 min. The TVSum contains a total of 50 videos, each video is annotated by 20 users, and the video duration lasts from 2 min to 10 min. Furthermore, we also augment the training videos with two extra datasets, OVP[19] and YouTube[19], containing 50 videos from YouTube and 39 videos, respectively. We follow previous works and report F-score under canonical (C), augmented (A), and transfer (T) settings.

Tab.1 shows the experimental results of different video summarization methods in the canonical setting.

Tab.1 Comparisons with the state-of-the-art approaches

Method	SumMe	TVSum	Average	Params
vsLSTM[7]	37.6	54.2	45.9	2.63
dppLSTM[7]	38.6	54.7	46.7	2.63
SUM-GAN[20]	41.7	56.3	49.0	295.86
DR-DSN[21]	42.1	58.1	50.1	2.63
FCSN[22]	48.8	58.4	53.6	36.58
CSNet[11]	48.6	58.5	53.6	100.76
SABTNet[12]	50.7	61.0	55.9	27.87
RSGN[23]	45.0	60.1	52.6	-
3DST-UNet[24]	47.4	58.3	52.9	-
Ours	54.4	60.1	57.3	12.09

It can be observed that the proposed architecture achieves a superior summarization performance on both datasets. Specifically, GDANet outperforms the state-of-the-art methods by at least 3.7% on more challenging SumMe and yields the best summarization performance. Although comparable or better F-scores belonging to SABTNet[12] and reconstructive sequence-graph network (RSGN)[23] are obtained on TVSum, our framework does not contain any recurrent architecture, thus having the ability to achieve parallel computing. Additionally, our method far surpasses vsLSTM[7], dppLSTM[7], DR-DSN[21] that simply rely on LSTM, which can be attributed to the fact that the self-attention mechanism is effective in modeling temporal cues. We list the number of parameters of various methods at the right of the table, from which we notice that compared to the models with good summarization performance, our approach contains fewer parameters, which indicates a significant balance between performance and parameters.

Tab.2 presents the experimental results of the experiments in augmented and transfer settings. Compared with the results in the canonical setting, the performance under the augmented setting on both SumMe and TVSum datasets has been further improved because the augmented dataset can provide more training data and manual annotations, effectively suppressing the over-fitting. Obviously, the F-score under the transfer setting is lower than that in both canonical and augmented settings since it is difficult to train a model by using datasets from different domains.

Tab.2 Comparisons with the state-of-the-art approaches under three evaluation settings

Method	SumMe			TVSum		
	C	A	T	C	A	T
vsLSTM[7]	37.6	41.6	40.7	54.2	57.9	56.9
dppLSTM[7]	38.6	42.9	41.8	54.7	59.6	58.7
SUM-GAN[20]	41.7	43.6	-	56.3	61.2	-
DR-DSN[21]	42.1	43.9	42.6	58.1	59.8	58.9
FCSN[22]	48.8	50.2	45.0	58.4	59.1	57.4
CSNet[11]	48.6	48.7	44.1	58.5	57.1	57.4
RSGN[23]	45.0	45.7	44.0	60.1	61.1	60.0
3DST-UNet[24]	47.4	49.9	47.9	58.3	58.9	56.1
Ours	54.4	54.9	46.9	60.1	60.3	57.2

To analyze the temporal modeling capability of different networks, we have conducted experiments by replacing self-attention with Bi-LSTM and Bi-GRU. During the experiments, we set the hidden layer dimension of RNNs to 512 to obtain the same size of contextual features as ours. The experimental results on both SumMe and TVSum are shown in Tab.3. Compared to the above RNNs-based cases, the performance of our summarization framework achieves a better F-score on both SumMe and TVSum, which can demonstrate the effectiveness of self-attention.

Tab.3 Experimental results of different methods of capturing contextual information

Method	SumMe			TVSum		
	C	A	T	C	A	T
Bi-LSTM	53.1	54.3	48.6	58.0	59.2	53.5
Bi-GRU	53.3	54.3	47.6	58.5	59.5	55.5
Ours	54.4	54.9	46.9	60.1	60.3	57.2

We have conducted parameter analysis by setting a different number of attention heads in global and local attention pathways, and the result is presented in Fig.6. We can observe that the number of attention heads is

critical to the performance of the model since multi-head attention can capture rich feature information in different representation subspaces. When the number of global and local attention heads is set to 2 and 4, the best performance is achieved on both SumMe and TVSum datasets, reaching 54.4% and 60.1%, respectively. This is because employing a proper number of attention heads can not only effectively capture the context information of several aspects, but also avoid the problems caused by over-fitting.

Tab.4 explores the ablation results of global and local contextual information, from which we can observe that summarization performance decreases once lacking global or local contextual information. More particularly, the model without global context degrades F-score performance more significantly than that without local context. It can be attributed to the fact that humans can better locate key shots after watching the entire video. When both global and local contextual information is considered, the performance achieves the best.

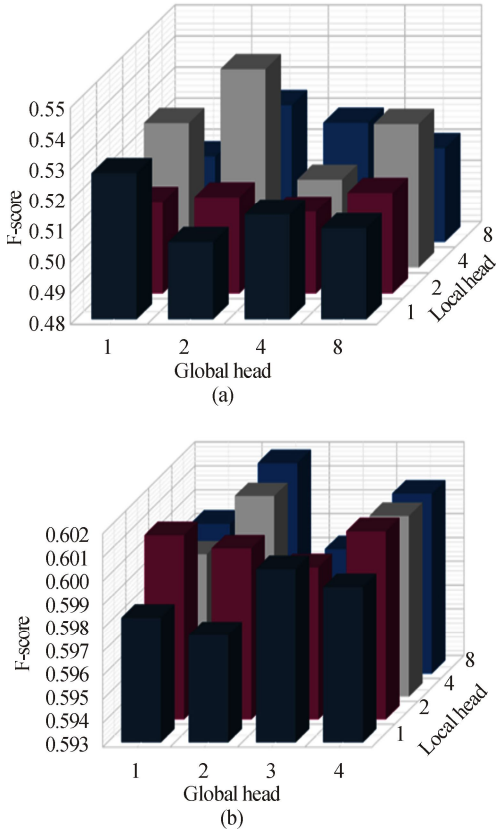


Fig.6 Sensitivity analysis of the number of attention heads: (a) F-scores on SumMe; (b) F-scores on TVSum

Tab.4 Ablation results of global and local contextual information

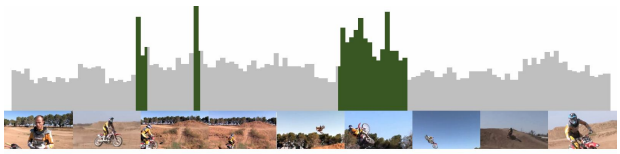
Method	SumMe	TVSum
GDANet w/o local	53.9	59.9
GDANet w/o global	52.4	58.7
GDANet	54.4	60.1

We conduct ablation experiments to verify the effectiveness of different modules by gradually appending them to the network in Tab.5. The model (Example 2) combining DSAM outperforms our baseline (Example 1), which is only composed of an importance score predictor. F-scores are improved by 1.8% on SumMe and 2.4% on TVSum, respectively. This can be explained by the fact that contextual information is important to understand video content. Additionally, the most significant is that DOM (Example 3) is capable of further enhancing the F-score performance by 3.1% and 0.9% on both datasets, which may benefit from the discriminative features learned by our carefully designed feature enhancement approach. It is worth noting that we adopt element-wise addition for feature fusion towards global and local contextual information. However, it may confront insufficiency during fusing features with a significant semantic difference. The summarization performance is improved by 2.3% and 0.1% on SumMe and TVSum once AFFM is employed. The empirical results shown above demonstrate the effectiveness of the proposed summarization framework.

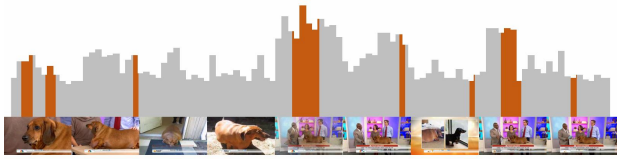
To more intuitively validate the summary performance, we provide visualization results. As shown in Fig.7, the light-colored bars represent the true importance scores, and the gray and different-colored bars stand for manual summaries and predicted summaries by our model. Moreover, some sample frames are also presented at the bottom.

Tab.5 Ablation study of different modules

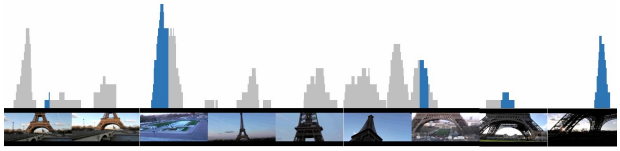
Exp. No	DSAM	DOM	AFFM	SumMe	TVSum
1				47.2	56.7
2	√			49.0	59.1
3	√	√		52.1	60.0
4	√	√	√	54.4	60.1



(a) Example 1



(b) Example 2



(c) Example 3

Fig.7 Visualization results of the proposed approach

We use three videos from the SumMe and TVSum datasets to demonstrate the experimental results. These videos concern sports, pets, and buildings, respectively. It can be clearly observed that GDANet is inclined to select key shots with higher importance scores. These generated summaries include less un-important and meaningless content while including various visual contents from which the viewers can infer the storyline.

In this paper, we have proposed a GDANet, aiming at accurately summarizing videos by enhancing features. Firstly, DOM is proposed, which exploits the semantic difference between each frame and video to promote discriminative feature learning. Then we propose DSAM to capture both coarse and fine-grained contextual information. Finally, we devise AFFM to effectively incorporate the global and local contexts into a concise feature representation, essentially and comprehensively revealing the video content. The summaries are generated by predicting the importance scores. We have carried out extensive experiments on benchmark datasets, and the experimental results demonstrate the effectiveness of the proposed framework. In the future, we will attempt to achieve a query-focused summarization method to generate summaries consistent with viewers' preferences.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] APOSTOLIDIS E, ADAMANTIDOU E, METSAI A I, et al. Video summarization using deep neural networks: a survey[J]. Proceedings of the IEEE, 2021, 109(11): 1838-1863.

[2] LEI J, LUAN Q, SONG X, et al. Action parsing-driven video summarization based on reinforcement learning[J]. IEEE transactions on circuits and systems for video technology, 2018, 29(7): 2126-2137.

[3] HUANG C, WANG H. A novel key-frames selection framework for comprehensive video summarization[J]. IEEE transactions on circuits and systems for video technology, 2019, 30(2): 577-589.

[4] YUAN L, TAY F E H, LI P, et al. Cycle-SUM: cycle-consistent adversarial LSTM networks for unsupervised video summarization[C]//Proceedings of the AAAI Conference on Artificial Intelligence, January 27-February 1, 2019, Hawaii, USA. Washington: AAAI, 2019, 33(01): 9143-9150.

[5] CHU W S, SONG Y, JAIMES A. Video co-summarization: video summarization by visual co-occurrence[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, USA. New York: IEEE, 2015: 3584-3592.

[6] MEI S, GUAN G, WANG Z, et al. $L_{2,0}$ constrained sparse dictionary selection for video summarization[C]//2014 IEEE International Conference on Multimedia and Expo, July 14-18, 2014, Chengdu, China. New York: IEEE, 2014: 1-6.

[7] ZHANG K, CHAO W L, SHA F, et al. Video summarization with long short-term memory[C]//European Conference on Computer Vision, October 10-16, 2016, Amsterdam, Netherlands. Berlin: Springer, 2016: 766-782.

[8] YUE-HEI N G J, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: deep networks for video classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, USA. New York: IEEE, 2015: 4694-4702.

[9] ZHAO B, LI X, LU X. Hierarchical recurrent neural network for video summarization[C]//Proceedings of the 25th ACM International Conference on Multimedia, October 23-27, 2017, Orlando, USA. New York: ACM, 2017: 863-871.

[10] ZHAO B, LI X, LU X. HSA-RNN: hierarchical structure-adaptive RNN for video summarization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, USA. New York: IEEE, 2018: 7405-7414.

[11] JUNG Y, CHO D, KIM D, et al. Discriminative feature learning for unsupervised video summarization[C]//Proceedings of the AAAI Conference on Artificial Intelligence, January 27-February 1, 2019, Hawaii, USA. Washington: AAAI, 2019, 33(01): 8537-8544.

[12] FU H, WANG H. Self-attention binary neural tree for video summarization[J]. Pattern recognition letters, 2021, 143: 19-26.

[13] KANAFANI H, GHAURI J A, HAKIMOV S, et al. Unsupervised video summarization via multi-source features[C]//Proceedings of the 2021 International Conference on Multimedia Retrieval, November 16-19, 2021. New York: ACM, 2021: 466-470.

[14] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, USA. New York: IEEE, 2015: 1-9.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[16] POTAPOV D, DOUZE M, HARCHAOUI Z, et al. Category-specific video summarization[C]//European Conference on Computer Vision, September 5-12, 2014, Zurich, Switzerlan. Berlin: Springer, 2014: 540-555.

[17] GYGLI M, GRABNER H, RIEMENSCHNEIDER H, et al. Creating summaries from user videos[C]//European Conference on Computer Vision, September 5-12, 2014, Zurich, Switzerlan. Berlin: Springer, 2014: 505-520.

[18] SONG Y, VALLMITJANA J, STENT A, et al. TVSUM: summarizing web videos using titles[C]//

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, USA. New York：IEEE, 2015：5179-5187.

[19] DE AVILA S E F, LOPES A P B, DA LUZ J R A, et al. VSUMM：a mechanism designed to produce static video summaries and a novel evaluation method[J]. Pattern recognition letters, 2011, 32(1)：56-68.

[20] MAHASSENI B, LAM M, TODOROVIC S. Unsupervised video summarization with adversarial LSTM networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Hawaii, USA. New York：IEEE, 2017：202-211.

[21] ZHOU K, QIAO Y, XIANG T. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward[C]//Proceedings of the AAAI Conference on Artificial Intelligence, February 2-7, 2018, New Orleans, USA. Washington：AAAI, 2018, 32(1).

[22] ROCHAN M, YE L, WANG Y. Video summarization using fully convolutional sequence networks[C]//Proceedings of the European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Berlin：Springer, 2018：347-363.

[23] ZHAO B, LI H, LU X, et al. Reconstructive sequence-graph network for video summarization[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(5)：2793-2801.

[24] LIU T, MENG Q, HUANG J J, et al. Video summarization through reinforcement learning with a 3D spatio-temporal U-Net[J]. IEEE transactions on image processing, 2022, 31：1573-1586.