AudioVisual Video Summarization

Bin Zhao, *Member, IEEE*, Maoguo Gong, *Senior Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*— Audio and vision are two main modalities in video data. Multimodal learning, especially for audiovisual learning, has drawn considerable attention recently, which can boost the performance of various computer vision tasks. However, in video summarization, most existing approaches just exploit the visual information while neglecting the audio information. In this brief, we argue that the audio modality can assist vision modality to better understand the video content and structure and further benefit the summarization process. Motivated by this, we propose to jointly exploit the audio and visual information for the video summarization task and develop an audiovisual recurrent network (AVRN) to achieve this. Specifically, the proposed AVRN can be separated into three parts: 1) the two-stream long-short term memory (LSTM) is used to encode the audio and visual feature sequentially by capturing their temporal dependency; 2) the audiovisual fusion LSTM is used to fuse the two modalities by exploring the latent consistency between them; and 3) the self-attention video encoder is adopted to capture the global dependency in the video. Finally, the fused audiovisual information and the integrated temporal and global dependencies are jointly used to predict the video summary. Practically, the experimental results on the two benchmarks, i.e., SumMe and TVsum, have demonstrated the effectiveness of each part and the superiority of AVRN compared with those approaches just exploiting visual information for video summarization.**

***Index Terms*— Audiovisual learning, multimodal learning, recurrent network, video summarization.**

I. Introduction

Video summarization is a typical computer vision task developed for video analysis. It can distill the video information effectively by extracting several key-frames or key-shots to display the video content [1]. Under the help of video summary, the viewer can perceive the information without watching the whole video. Therefore, video summary provides an efficient way for video browsing. Moreover, by removing the redundant and meaningless video content, it has potential applications in video retrieval, storage, and indexing [2], [3], as well as boosting the performance of related video analysis tasks, such as video captioning [4] and action recognition [5].

In the data explosion era, video summarization draws increasing attention. Lots of approaches are proposed in recent years [6]–[8]. The existing approaches mainly summarize videos in two aspects. On one hand, comprehensive models are developed to summarize videos according to manual criteria [9]–[11], including representativeness, importance, interestingness, and so on. They are developed to select the key-frames or key-shots that represent the whole video content, contain the important objects, have less redundancy, etc., so as to distill the video information effectively. On the other hand, the video data are taken as a sequence of frames. The summary is generated according to the temporal dependencies among frames. To achieve this, the most popular recurrent neural network, long-short term memory (LSTM) [12], is used as the backbone in video summarization. Recently, various sequence models are developed based on it, such as bidirectional LSTM [13], hierarchical LSTM [14], [15], and attention-based LSTM [16], [17]. By taking advantage of the deep learning and sequential modeling ability of LSTM, they surpass the traditional approaches developed based on manual criteria and take the leading position.

A. Motivation and Overview

Video data are naturally composed of two modalities, i.e., audio and vision. They record the activities from different aspects and cooperate together to help the viewer understand the video content. Recently, multimodal learning has proved that audio and vision modalities share a consistency space, and there are semantic relationships between them [18], [19]. Lots of relevant video analysis tasks have demonstrated that the performance is promoted using the multimodal information in previous single modality tasks [20]–[22].

However, few researchers in video summarization have recognized the potential contributions of audio information to the performance. Most of them just consider the vision modality and extract shallow or deep visual features to represent video frames, while the audio features are ignored.

In this brief, we argue that the audio modality can assist the vision modality to better understand the video content and structure. Concretely, the audio and vision are complementary to present activities in different modalities. For example, the music at the party reflects the pleasant atmosphere of the scene, and the cheers in soccer games indicate a good goal. However, the audiovisual inconsistency situations usually occur in videos as well. For example, the sounding object is not in the field of view. It will also bring interferences for vision modality, which is the main challenge in audiovisual video summarization.

Facing the above opportunities and challenges, we propose an audiovisual recurrent network (AVRN) to jointly use audio and visual information in the video summarization task. To guarantee the consistency, the fusion of audio and visual information is in two stages. The two-stream LSTM is used in the first stage to encode the audio and visual features sequentially and capture their temporal dependency. Then, the audiovisual fusion LSTM is developed to exploit the consistency space between audio and visual information and fuse them with an adaptive gating mechanism. Besides, considering that there are multihop storylines in the video stream due to montage and edit, the temporal neighborhood dependency is not capable enough for video summarization. In this case, a self-attention video encoder is adopted to encode the global video information. Finally, the temporal and global dependencies of audiovisual information captured by the sequence encoder and global encoder are jointly used for predicting the video summary.

Manuscript received January 18, 2021; revised July 10, 2021; accepted October 4, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102200, in part by the National Natural Science Foundation of China under Grant 62106183, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2021JQ-204, and in part by the China Postdoctoral Science Foundation under Grant 2020TQ0236. *(Corresponding author: Bin Zhao.)*

Bin Zhao is with the Academy of Advanced Interdisciplinary Research, Xidian University, Xi'an 710071, China, and also with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: binzhao111@gmail.com).

Maoguo Gong is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Xidian University, Xi'an 710071, China (e-mail: gong@ieee.org).

Xuelong Li is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xuelong_li@nwpu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3119969.

Digital Object Identifier 10.1109/TNNLS.2021.3119969

B. Novelties and Contributions

The novelties and contributions of our work are as follows.

1) The audio information is introduced to the video summarization task, so as to complement with the visual information in video content and structure modeling.
2) A hierarchical multimodal LSTM is developed to exploit the latent consistency space between the two modalities and capture the temporal dependencies.
3) By combining the LSTM and self-attention encoder together, the global and temporal dependencies are captured jointly to benefit the summarization process.

C. Organization

The organization of the rest brief is presented as follows. The relevant works in the literature are analyzed in Section II. The proposed AVRN is described in Section III. The experimental results are discussed and compared with the state-of-the-arts in Section IV. Finally, the conclusion of our work is drawn in Section V.

II. Related Works

A. Traditional Video Summarization

Earlier works devote to find a set of representatives to summarize the video content. Hand-crafted feature extractors are first used to extract feature vectors for each frame, such as color histogram [23], optical flow [24], and histograms of gradient [25]. To determine the representativeness, clustering algorithms and dictionary learning methods are used to generate the video summary [26], [27]. For example, k-means [28] and k-medoids [29] allocate frames into different clusters and obtain the cluster center. Naturally, the cluster centers are viewed as the representatives and selected in the summary. Dictionary learning is also an effective method to select the representatives. It takes the frame sequence as a dictionary and tries to determine a subset of the elements to represent the original video [30], [31]. Sparsity is a widely used prior for dictionary-based video summarization. To achieve this, the l_0 and $l_{0,1}$ norms are added as the regularizer, and the block sparsity constraint is designed to speed up the convergence [32]–[34].

Later, researchers have realized that the representativeness is not enough to quantify the summary quality, and more manual criteria are proposed. A user attention model is constructed in [35] by combining the visual, audio, and textual information. It shows the superiority of multimodal information and inspires our work. Diversity is designed to reduce the redundancy in the summary, where similar clips are removed from the video. In this case, the key point is to determine the similarity among frames and shots. In [36], Segeral distance is used as the similarity metric. Furthermore, several distance metrics are combined together in [37], and the determinantal point process (DPP) model is modified to select the diverse key-frames sequentially [38]. Importance is developed to constrain the summary to maintain important objects in the video, in which several local features are used to determine the importance of different objects [39], [40], including distance to the frame centroid, frequency of occurrence, esthetics metrics, etc. To model the summary comprehensively, several criteria are combined together in [24] and [41], including importance, representativeness, uniformity, and storyness. They measure the summary quality in different aspects and provide a comprehensive score function.

B. Deep-Learning-Based Video Summarization

Recently, deep-learning-based approaches have made tremendous progress and taken the leading position in video summarization [8], [42]. They prefer to learn the complex mapping from video to summary directly by taking advantage of the learning ability of deep networks. Specifically, RNNs are used to model the frame sequence. In [13], the bidirectional LSTM is developed to exploit the temporal dependency. Considering that the most favorable video length for LSTM is less than 100 frames, a stacked memory network is proposed in [43], where the LSTM is augmented with a memory module to boost the performance on long-term dependency modeling. Similarly, a hierarchical LSTM is developed to extend the capability of dealing with long sequence [14]. Different from recursive models, a self-attention model is adopted in [17] as the video encoder, and a fully convolutional sequence network is conducted in [44]. However, they both just consider one side of the dependencies among frames, i.e., temporal dependencies or global dependencies, while neglecting the other one.

To boost the performance, different mechanisms are developed to optimize the sequence model [6], [7], [45]. An attentive and semantic preserving model is proposed in [42] to explore the inherent relationships and minimize the semantic information loss between video and summary. A dual mixture attention is developed in [46], which shows better generalization to small datasets. In SUM-GAN [47], a discriminator is conducted after the summary generator, and the generated summary is discriminated by the adversarial loss. The encoder–decoder architecture is designed in [48], where the decoder is used to recover the video content from the obtained summary in the semantic space. It can guide the encoder to select key-shots that best represent the video content. Similarly, a dual learning framework is developed in [1] based on a summary generator and a video reconstructor. In this case, the unsupervised optimization of the summary generator is achieved. The manual criteria are also adopted in [49], and the reinforcement learning strategy is adopted to reward the summary generator.

C. Multimodal Video Analysis

Recently, researchers have realized the multimodal characteristics of video data and developed lots of interesting tasks for video analysis. A spatial–temporal graph is proposed for the video captioning task in [50], which can translate the visual modality into text modality. A transformer with instance attention is conducted in [20] to jointly use audiovisual information for event localization. A two-stream architecture is developed based on the transformer to fuse text and visual information for video retrieval [51]. Furthermore, there are also lots of works developed for the multimodal representation learning [52], cross-modal consistency learning [53], multimodal generation tasks [18], etc. All of them inspire us to develop a multimodal work for video summarization.

III. Proposed Approach

In this brief, we propose an AVRN to integrate the audio and visual information together to the video summarization task. As depicted in Fig. 1, the proposed AVRN is composed of three parts, where the two-stream LSTM encodes the audio and visual feature sequentially, the fusion LSTM fuses the audiovisual multimodal information dynamically, and the self-attention module captures the video information globally. In the following, they are elaborated successively.

A. Two-Stream LSTM

LSTM is a typical recurrent neural network developed to deal with sequence data. Videos are naturally temporal sequence data. To encode the audio and visual information, the video data are separated into the audio signal and visual frames, and a two-stream structure of LSTM is conducted.

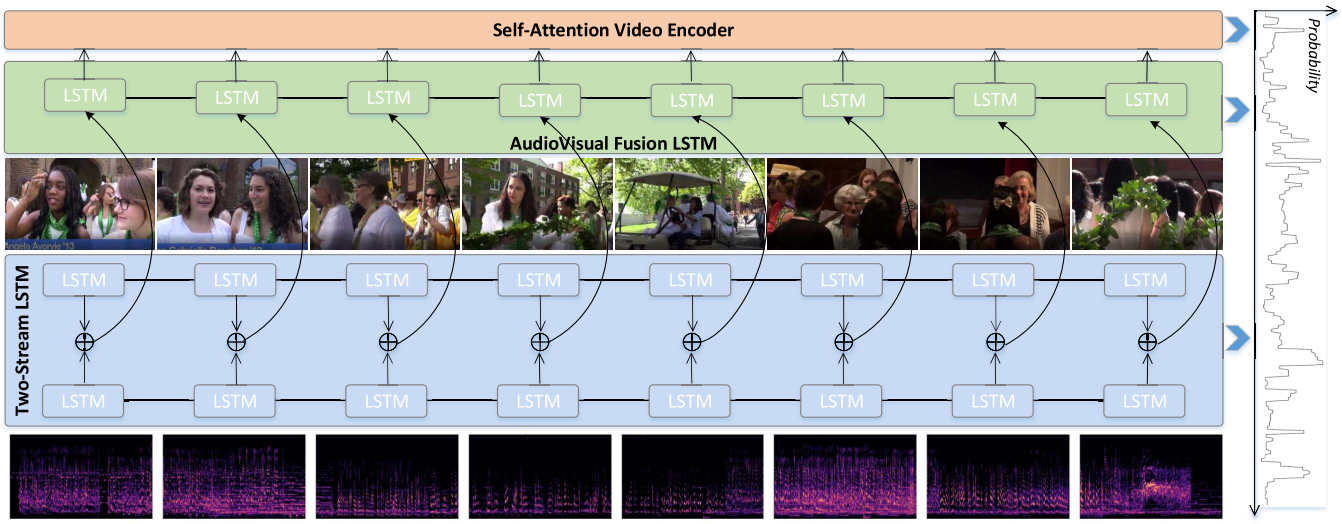

Fig. 1. Architecture of the proposed AVRN, which is composed of the two-stream LSTM, the audiovisual fusion LSTM, and the self-attention video encoder. The last row displays the log-mel spectrograms of audio data.

Specifically, given the frame sequence, the visual features are first extracted as $\mathbf{X}^v = \{\mathbf{x}_1^v, \mathbf{x}_2^v, \ldots, \mathbf{x}_n^v\}$, where n stands for the length of the frame sequence. Then, to capture the temporal dependencies among frames, the bidirectional LSTM is used as one stream to process $\mathbf{X}^v$ sequentially, which is formulated as

$$\mathbf{h}_t^v = \text{BiLSTM}\left(\mathbf{x}_t^v, \mathbf{h}_{t-1}^v\right) \tag{1}$$

where $\mathbf{h}_t^v$ is the hidden state of the bidirectional LSTM. Practically, BiLSTM is conducted by combining two LSTMs together. They capture the temporal dependency among frames from the forward and reverse directions, respectively, and encode the dependency into the hidden state vector $\mathbf{h}_t^v$ in each step.

Similarly, to exploit the temporal dependency of audio features $\mathbf{X}^a = \{\mathbf{x}_1^a, \mathbf{x}_2^a, \ldots, \mathbf{x}_n^a\}$, another stream of bidirectional LSTM is conducted, i.e.,

$$\mathbf{h}_t^a = \text{BiLSTM}\left(\mathbf{x}_t^a, \mathbf{h}_{t-1}^a\right). \tag{2}$$

In the two-stream LSTM, the temporal dependencies among audio and visual modalities are encoded into $\mathbf{h}_t^a$ and $\mathbf{h}_t^v$ separately. They are not fused correctly, and the difference and consistency between them are not considered, which may cause interference to the video summarization task. To address this problem, an audiovisual fusion LSTM is further developed.

B. Audiovisual Fusion LSTM

In this part, an audiovisual fusion LSTM is conducted to explore the sharing latent space among the audio and visual modality, so as to exploit the consistency and reduce the difference. Different from the two-stream structure, the audiovisual fusion LSTM takes the combined audio and visual information as input and fuses them sequentially.

Specifically, an adaptive gating mechanism is adopted to the audiovisual fusion LSTM, which is formulated as

$$c_t = \text{Sigmoid}\left(\mathbf{W}^a \mathbf{h}_t^a + \mathbf{W}^v \mathbf{h}_t^v + b\right) \tag{3}$$

where $\mathbf{W}^a$, $\mathbf{W}^v$, and b are the training parameters. c_t is the gate to control the information flow of different modalities, which is operated as follows:

$$\mathbf{x}_t^{\text{av}} = c_t \mathbf{h}_t^a + (1 - c_t)\mathbf{h}_t^v. \tag{4}$$

Then, the fused audio and visual information $\mathbf{x}_t^{\text{av}}$ is input to the audiovisual fusion LSTM to encode them sequentially, i.e.,

$$\mathbf{h}_t^{\text{av}} = \text{BiLSTM}\left(\mathbf{x}_t^{\text{av}}, \mathbf{h}_{t-1}^{\text{av}}\right). \tag{5}$$

Practically, $\mathbf{h}_t^{\text{av}}$ captures the temporal dependencies of the audiovisual information at the tth step, which is essential for summary generation.

C. Self-Attention Video Encoder

Although videos are naturally sequential data, multihops of storyline usually occur in the video stream, and the activities recorded in each shot vary largely. In some occasions, there are no obvious temporal dependencies among consecutive shots, such as those videos with edit and montage. Therefore, the sequence networks are not enough for modeling the complex video structure and content.

In this case, we develop a global encoder to cooperate with the sequence model. It is achieved by a self-attention module

$$\mathbf{V}_t = \sum_{i=1}^{n} \alpha_t^i \mathbf{x}_i^{\text{av}} \tag{6}$$

where $\mathbf{V}_t$ is the encoded global dependency among frames at the tth step. α_t^i is the attention weight, which is computed by

$$l_t^i = \left\langle \mathbf{W}^1 \mathbf{x}_i^{\text{av}}, \mathbf{W}^2 \mathbf{x}_t^{\text{av}} \right\rangle \tag{7}$$

$$\alpha_t^i = \exp\left\{l_t^i\right\} \Big/ \sum_{i=1}^{n} \exp\left\{l_t^i\right\} \tag{8}$$

where $\mathbf{W}^1$ and $\mathbf{W}^2$ are the training weights. $\langle \cdot, \cdot \rangle$ denotes the inner-product operation. l_t^i captures the dependency between $\mathbf{x}_i^{\text{av}}$ and $\mathbf{x}_t^{\text{av}}$. Equation (8) is used to normalize the attention weight.

D. Summary Generation

Given the computed temporal and global dependencies of audiovisual information, the importance of each frame to the video content is computed by

$$p_t = \text{Sigmoid}\left(\mathbf{W}^p \left[\mathbf{x}_t^{\text{av}}, \mathbf{h}_t^{\text{av}}, \mathbf{V}_t\right] + b^p\right) \tag{9}$$

where $\mathbf{W}^p$ and b^p are the training weight and bias, respectively. $\mathbf{x}_t^{\text{av}}$, $\mathbf{h}_t^{\text{av}}$, and $\mathbf{V}_t$ denote the fused audiovisual information, the temporal

dependency, and the global video information, respectively. They are integrated together to predict the frame-level importance. Then, the shot-level importance is obtained by averaging the importance scores of those frames in the shot, i.e.,

$$p_i^s = \sum_{t=s_{i-1}+1}^{t=s_i} p_t \tag{10}$$

where $\mathbf{S} = [0, s_1, s_2, \ldots, s_{m-1}, n]$ are the shot boundaries of m shots in the video. Following the existing protocols, they are computed by the kernel temporal segmentation (KTS) method [9]. In the final, the video summary is generated with those higher score shots.

Practically, the proposed AVRN is trained end to end under the supervision from human-created summaries. The mean square error (MSE) is used for optimization, i.e.,

$$\text{loss} = \frac{1}{n}\|\mathbf{p} - \mathbf{g}\|_2^2 \tag{11}$$

where $\mathbf{p}$ is the predicted frame-level importance vector, and $\mathbf{g}$ is the ground truth annotated by human beings.

The proposed AVRN is operated on Python 3.7 with the deep learning platform of PyTorch 1.6. The dimensionalities of the hidden states in both the two-stream LSTM and audiovisual fusion LSTM are fixed as 256. The optimizer is selected as Adam with the learning rate 1e−5, the decay rate 0.1, and decay step 30. Generally, AVRN can reach the convergence in less than 60 epochs.

IV. Experiments

The experiments are carried out on two benchmark datasets, SumMe [25] and TVsum [54]. In the following subsections, the ablation studies are conducted, and several state-of-the-art approaches are compared.

A. Setup

1) Dataset Introduction: SumMe and TVsum are popular video summarization datasets. The SumMe dataset consists of 25 videos with diverse topics, including cooking, traveling, sports, etc. Most of them are raw videos without human edit. For each video, 15–18 users are used to select key-shots and generate the video summaries. The TVsum dataset is composed of 50 videos in open domain. They are edited videos about news, cooking, pets, etc. Each of them contains 20 annotations of shot-level importance scores, where the shots are generated by segmenting the video into 2-s clips evenly.

Following the existing protocols, the videos in SumMe and TVsum are separated into 80% for training and the rest 20% for testing. For simplicity, the validation is operated on the training set. In the training process, the human-created summaries in SumMe and TVsum are modified to the frame-level importance scores, which are taken as the supervision information for AVRN optimization. Practically, the summarization performance varies among different videos. To make the results more convincing, random training/test splits are carried out for five times, and the average performance is taken as the final results.

Besides, the OVP [28] and YouTube [28] datasets are used in the experimental part to augment the training set. They are composed of 89 videos with human-created summaries.

2) Feature Extraction: For visual information, GoogLeNet [55] pretrained on ImageNet[1] is adopted to extract visual features for video frames, and the 1024-dim feature vector in pool5 layer is taken as the frame representation. Particularly, considering that neighboring frames are quite similar to each other, the features are extracted for every 15 frames (about 2 fps).

For audio information, VGGish [56] pretrained on Audioset[2] is adopted for audio feature extraction. Specifically, the audio data are temporally separated with a duration of 1 s. Two neighboring segments share the overlap of half a second, so that the length of audio feature can match that of the visual feature. Note that there are two videos in SumMe without audio data, "Scuba" and "St Maarten Landing." Their audio features are padded with zeros.

3) Performance Evaluation: The summary quality is evaluated by measuring the temporal consistency between the predicted summary and human-created summary. Precision (P), recall (R), and F-measure (F) are widely adopted metrics. They are defined as

$$P = \frac{\#\left(\text{summary}^p \cap \text{summary}^h\right)}{\#\text{summary}^p}$$
$$R = \frac{\#\left(\text{summary}^p \cap \text{summary}^h\right)}{\#\text{summary}^h}$$
$$F = \frac{P \cdot R}{(P + R)}$$

where $\#\text{summary}^p$ and $\#\text{summary}^h$ denote the number of frames in the predicted summary and human-created summary, respectively. $\#(\text{summary}^p \cap \text{summary}^h)$ stands for their overlap. Considering that each video contains multiple human-created summaries, the pairwise comparisons are conducted. Following the existing protocols, the maximum scores are taken as the results on SumMe. The average scores are taken as the results on the TVsum dataset.

To evaluate the performance comprehensively, the rank-based evaluation metrics are also adopted in this work. They are Kendall's τ and Spearman's ρ [57], which measure the correlation coefficients of the generated importance scores and annotated importance scores. The pairwise evaluations are conducted among multiple annotated importance scores, and the average coefficients are taken as the final results on both the SumMe and TVsum datasets.

B. Ablation Studies

The proposed AVRN is composed of three parts, including the two-stream LSTM (TS-LSTM), audiovisual fusion LSTM (AVF-LSTM), and the self-attention video encoder (SAVE). To verify the effectiveness of each part, the ablation studies are conducted, and several baselines are compared, including the following.

1) Audio LSTM: A bidirectional LSTM is used to encode the audio feature and generate the video summary, while the visual feature is ignored.
2) Visual LSTM: A bidirectional LSTM is used to encode the visual feature and generate the video summary, while the audio feature is ignored.
3) TS-LSTM: Only the two-stream LSTM is conducted to capture the temporal dependency among audio and visual features and generate the video summary.
4) AVF-LSTM: The audio and visual features are fused directly without exploiting their temporal dependencies.
5) AVRN(w/o SAVE): The global video information is not considered when predicting the video summary.
6) AVRN(single): The bidirectional LSTMs in TS-LSTM and AVF-LSTM are replaced with single LSTM (forward LSTM).

Table I exhibits the results of ablation studies. The audio LSTM can get satisfactory results just with the audio feature as input, which indicates that the audio modality can indeed provide useful information for video summarization. TS-LSTM outperforms the visual LSTM and audio LSTM. It proves the necessity to integrate

[1] http://www.image-net.org/

[2] https://research.google.com/audioset/index.html

TABLE I
RESULTS OF ABLATION STUDIES FOR AVRN

Datasets	SumMe			TVsum		
Metrics	Precision	Recall	F-measure	Precision	Recall	F-measure
Audio LSTM (audio)	0.339	0.380	0.354	0.538	0.540	0.539
Visual LSTM (visual)	0.376	0.411	0.386	0.554	0.553	0.554
TS-LSTM (audio&visual)	0.383	0.419	0.394	0.577	0.578	0.578
AVF-LSTM (audio&visual)	0.389	0.417	0.396	0.567	0.568	0.567
AVRN(w/o SAVE) (audio&visual)	0.395	0.439	0.415	0.583	0.583	0.583
AVRN(single) (audio&visual)	0.414	0.435	0.419	0.589	0.587	0.589
AVRN (audio&visual)	0.428	0.464	0.441	0.598	0.598	0.597

audio and visual information together to boost the performance. AVF-LSTM takes the audio and visual features as input and fuses them without using the TS-LSTM to capture their temporal dependency. AVRN(w/o SAVE) equals to the combination of TS-LSTM and AVF-LSTM. AVRN(w/o SAVE) outperforms them. It verifies the importance of temporal dependency among audio and visual features and the necessity to fuse the audio and visual information to achieve the mutual benefit between them.

The difference between the proposed AVRN and the baseline AVRN(w/o SAVE) lies in that the self-attention video encoder is not included in AVRN(w/o SAVE). It means that only the temporal dependency is captured in AVRN(w/o SAVE), while the global video information is ignored. The better performance of the proposed AVRN indicates that the global video information are also very important to the summary quality. Besides, AVRN(single) gets worse results than AVRN that uses bidirectional LSTMs to encode and fuse the audiovisual information. It explains the rationality to capture the bidirectional temporal dependencies jointly for video summarization. Overall, the results in Table I have demonstrated the effectiveness of the proposed AVRN, including the two-stream LSTM, the audiovisual fusion LSTM, and the self-attention video encoder.

C. Comparison With State-of-the-Arts

Table II presents the results of AVRN and traditional approaches. The compared approaches are in various categories. k-medoids, Delauny, and VSUMM are the clustering-based approaches. They are developed based on k-medoids, Delauny clustering, and k-means, respectively. Particularly, considering that the video frames vary smoothly, the clusters in VSUMM are initialized by segmenting video temporally, so that better results are achieved. It indicates that the domain knowledge of video is important for the summarization task. SALF, LiveLight and Block Sparse are the dictionary-learning-based approaches. SALF and Block Sparse generate summary by self-reconstructing the video with shots sparsely. LiveLight conducts an online learning strategy to incrementally select those shots that cannot be represented by current key-shot set. They get better performance than clustering-based approaches. It is mainly because they can capture the dependency among frames. CSUV and LSMO are designed based on manual criteria. CSUV selects key-shots according to their interestingness measured by factors such as esthetics, landmarks, faces, persons, and objects. Inspired by it, LSMO further constructs the interestingness, representativeness, and uniformity models. CSUV and LSMO are supervised approaches, so that they can outperform the unsupervised clustering and dictionary-learning-based approaches.

The proposed AVRN maintains the advantages of the traditional approaches. It can capture the temporal and global dependencies among frames by the LSTM and global video encoder. It can also be optimized under the supervision of human-created summaries. Moreover, AVRN uses audio and visual features jointly to predict the summary, which means more information is exploited than the traditional approaches just extracting visual features. Therefore, it outperforms the traditional approaches significantly.

TABLE II
RESULTS OF AVRN AND TRADITIONAL APPROACHES

Datasets	SumMe	TVsum
k-medoids [29]	0.334	0.288
Delauny [58]	0.315	0.394
VSUMM [28]	0.335	0.391
SALF [30]	0.378	0.420
LiveLight [59]	0.384	0.477
Block Sparse [34]	0.401	0.526
CSUV [25]	0.393	0.532
LSMO [60]	0.403	<u>0.568</u>
Summary Transfer [61]	<u>0.409</u>	–
AVRN	**0.441**	**0.597**

Table III presents the results of AVRN and other deep learning approaches, to show the superiority of AVRN. vsLSTM first uses bidirectional LSTM to capture the temporal dependency among frames and summarize the video. dppLSTM extends vsLSTM usin the DPP model to guarantee the diversity of key-shots. However, they are plain sequence models without considering the global video information, so they perform much worse than the proposed AVRN. vsLSTM-att and dppLSTM-att are modified by adding attention models to encode the global video information. By exploiting both the temporal and global video information, vsLSTM-att and dppLSTM-att perform much better than the original ones. A-AVS is also an attention-based LSTM model, which performs comparably with vsLSTM-att and dppLSTM-att. It has demonstrated the necessity to integrate global encoder to the sequence model for the video summarization task. That is why the self-attention video encoder is developed in AVRN. Furthermore, the even better performance of AVRN has verified the superiority of using both the audio and visual feature to summarize the video.

SUM-GAN proposes to use the generative adversarial network (GAN) to generate video summary discriminatively and uses the discriminator to achieve the unsupervised learning. SUM-GAN$_{\text{sup}}$ is its supervised version, which outperforms SUM-GAN considerably. SASUM and SASUM$_{\text{sup}}$ take the video captions as the auxiliary information to boost the performance. It is also a multimodal video summarization approach. However, the video captions require much human resources, which limits its applicability. Fortunately, the audio modality matches vision modality naturally in videos. Better results are obtained by AVRN, which shows the advantages of audiovisual fusion than text-visual fusion in the video summarization task. DR-DSN adopts the manual criteria, i.e., representativeness and diversity, to guide the summary generator. The results indicate that

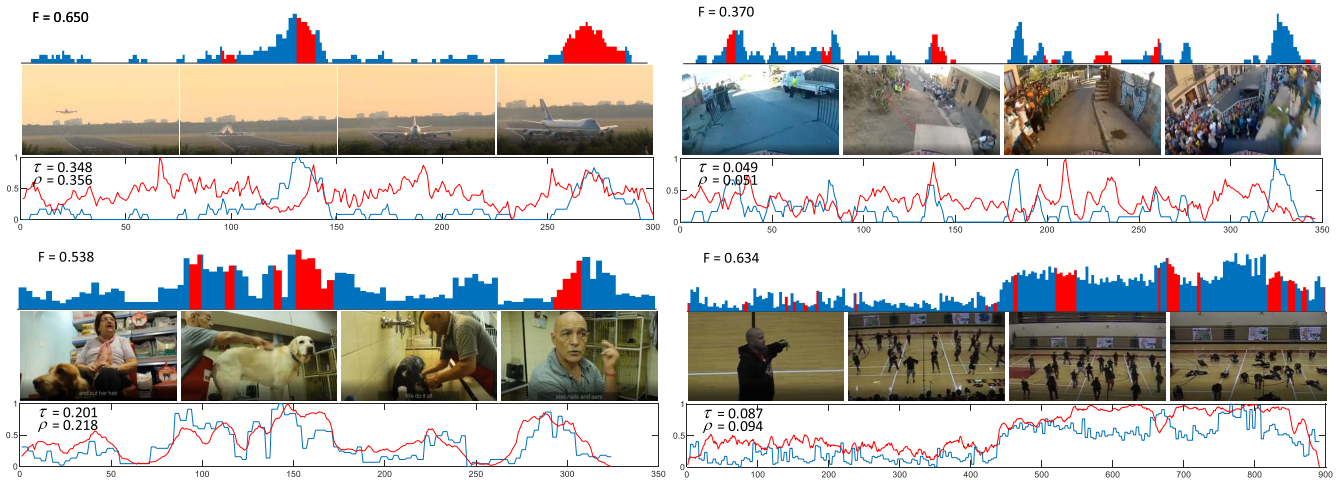

Fig. 2. Example summaries generated by AVRN. The above and below samples are from SumMe and TVsum, respectively. The blue curves and red curves below each sample are frame-level importance scores annotated by human and predicted by AVRN. The red histograms above each sample indicate the generated summary.

TABLE III
RESULTS OF AVRN AND DEEP LEARNING APPROACHES

Datasets	SumMe	TVsum
vsLSTM [13]	0.376	0.542
dppLSTM [13]	0.386	0.547
SUM-GAN [47]	0.387	0.508
SUM-GAN$_{sup}$ [47]	0.417	0.563
H-RNN [14]	0.421	0.579
HSA-RNN [15]	0.423	0.587
SASUM [62]	0.406	0.539
SASUM$_{sup}$ [62]	**0.453**	0.582
A-AVS [63]	0.439	0.594
DR-DSN [49]	0.414	0.576
DR-DSN$_{sup}$ [49]	0.421	0.581
SMIL [64]	0.412	0.513
vsLSTM-att [16]	0.432	–
dppLSTM-att [16]	0.438	–
WS-HRL [65]	0.436	0.584
AVRN	0.441	**0.597**

manual criteria in the traditional approaches can also promote the performance of deep learning approaches.

H-RNN and HSA-RNN are the hierarchical structures of LSTMs. They outperform those approaches with plain LSTM structures, such as vsLSTM and dppLSTM. It is mainly due to the better non-linear fitting ability of the hierarchical structure. Our AVRN also follows the hierarchical structure of LSTM, where the first layer is the two-stream LSTM and the second layer is the audiovisual fusion LSTM. They encode and fuse the audiovisual information hierarchically, so better performance are achieved than plain LSTMs. Furthermore, AVRN also surpasses H-RNN and HSA-RNN. It is mainly because AVRN extracts the multimodal features for video summarization, while H-RNN and HSA-RNN just extract visual features. AVRN can better understand the video content and structure with audiovisual information, so that better results are obtained than H-RNN and HSA-RNN.

In Table IV, different settings of training data are conducted to further analyze the results on the SumMe and TVsum datasets. They are canonical, augmented, and transfer, which are described as follows.

1) Canonical: The SumMe and TVsum datasets are trained and tested individually. The training/test splits are fixed as 80% and 20%.
2) Augmented: When training on the SumMe dataset, the videos in TVsum, OVP, and YouTube are used to augment the training set. Similar strategies are also conducted on the TVsum dataset.
3) Transfer: The videos in TVsum, OVP, and YouTube are used to train the SumMe dataset. Similarly, the videos in SumMe, OVP, and YouTube are used to train the TVsum dataset.

From Table IV, we can see that most of the approaches get better results under the augmented setting. This phenomenon indicates that the training data are not enough for most approaches, which leads to the overfitting problem. One effective way to address this problem is to provide more information for the summary generator. DR-DSN adopts the reinforcement learning scheme to exploit the summary properties to reward the summary generator, so better results are obtained than the plain LSTMs, including vsLSTM and dppLSTM. VASNet also conducts an attention model to select key-shots according to the encoded video information.[3] re-SEQ2SEQ develops a retrospective encoder to keep the consistency of the semantics between video and summary. It also develops a single LSTM as an extra to segment videos into shots. In contrast, the proposed AVRN follows a much compact end-to-end architecture. Overall, AVRN performs the best on the SumMe dataset and performs nearly the best on the TVsum dataset. It has verified the superiority of jointly using the audio and visual information in video summarization.

D. Evaluation on Rank-Based Metrics

Precision, recall, and F-measure quantify the summary quality by measuring the temporal consistency between the generated summary and human-created summary. They neglect more fine-grained human preference of video shots. To address this problem, the rank-based metrics, Kendall's τ and Spearman's ρ, are used in this part to provide comprehensive evaluation of summary quality. They measure the correlation between the predicted probability curve and human-annotated importance curve.

Table V presents the results evaluated on rank-based metrics. We can see that the result of the summary generated by random selection is zero. It means there is no correlation between randomly generated summary and human annotation. Besides, considering that each video contains multiple human annotations, the evaluation is also conducted among them via leave-one-out strategy. They get highest

[3]The results of VASNet are produced by modifying the released source code to the same experimental setting in this brief.

TABLE IV
RESULTS OF AVRN AND COMPARED APPROACHES IN DIFFERENT TRAINING DATA ORGANIZATIONS

Datasets	SumMe			TVsum		
Organizations	Canonical	Augmented	Transfer	Canonical	Augmented	Transfer
vsLSTM [13]	0.376	0.416	0.407	0.542	0.579	0.569
dppLSTM [13]	0.386	0.429	0.418	0.547	0.596	0.587
SUM-GAN [47]	0.387	0.417	–	0.508	0.589	–
SUM-GAN$_{sup}$ [47]	0.417	0.436	–	0.563	0.612	–
H-RNN [14]	0.421	0.438	–	0.579	0.619	**–**
HSA-RNN [15]	0.423	0.421	–	0.587	0.598	–
DR-DSN [49]	0.414	0.428	0.424	0.576	0.584	0.578
DR-DSN$_{sup}$ [49]	0.421	0.439	0.426	0.581	0.598	**0.589**
VASNet [17]	0.424	0.425	0.419	0.589	0.585	0.547
re-SEQ2SEQ [48]	0.425	**0.449**	**–**	**0.603**	**0.639**	**–**
AVRN	**0.441**	**0.449**	**0.432**	0.597	0.605	0.587

TABLE V
RESULTS OF RANK-BASED EVALUATION (KENDALL'S τ AND SPEARMAN'S ρ)

Datasets	SumMe		TVsum	
Metrics	Kendall's τ	Spearman's ρ	Kendall's τ	Spearman's ρ
Random selection	0.000	0.000	0.000	0.000
dppLSTM [13]	–	–	0.042	0.055
DR-DSN [49]	0.047	0.048	0.020	0.026
HSA-RNN [15]	0.064	0.066	0.082	0.088
AVRN	0.073	0.074	0.096	0.104
Human	0.205	0.213	0.177	0.204

scores in Table V, which shows there are considerable consistency among human annotations. The results meet our expectations.

Some typical RNN-based approaches are compared in Table V. dppLSTM is developed based on a plain bidirectional LSTM. DR-DSN conducts representativeness and diversity reward to the summary generator. HSA-RNN constructs the hierarchical structure of LSTM. The proposed AVRN surpasses most of them on Kendall's τ and Spearman's ρ. Besides, Fig. 2 displays some exemplar summaries generated by the proposed AVRN. It can be observed from them that AVRN is able to accurately predict the importance scores and effectively summarize the video. The results have demonstrated the superiority of AVRN: 1) the fusion of audio and visual features can provide more information for understanding the video content and structure, so as to benefit the video summarization process; 2) the hierarchical structure of LSTM can enhance the learning ability and further promote the performance; and 3) the temporal and global dependencies are both very important to the summarization task.

V. CONCLUSION

In this brief, we propose to introduce the audio information for the video summarization task and develop an AVRN to achieve the fusion of audiovisual features and boost the summarization performance. Specifically, AVRN contains three parts, including the two-stream LSTM, the audiovisual fusion LSTM, and the self-attention video encoder. Specifically, the two-stream LSTM can capture the temporal dependency among audio features and video features, respectively. The audiovisual fusion LSTM can exploit the latent consistency between audio and visual information. The self-attention video encoder can capture the global dependency in the whole video stream. The experimental results on SumMe and TVsum have demonstrated that: 1) the audiovisual multimodal feature can provide more information for the summarization task than the single visual feature; 2) the hierarchical structure can enhance the learning ability of LSTM; and 3) the fusion of audio and visual features and the integration of temporal and global dependencies are both necessary for the video summarization task.

REFERENCES

[1] B. Zhao, X. Li, and X. Lu, "Property-constrained dual learning for video summarization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3989–4000, Oct. 2020.
[2] L. Jin, Z. Li, and J. Tang, "Deep semantic multimodal hashing network for scalable image-text and video-text retrievals," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 5, 2020, doi: 10.1109/TNNLS.2020.2997020.
[3] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4147–4153.
[4] Y. Chen, S. Wang, W. Zhang, and Q. Huang, "Less is more: Picking informative frames for video captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 367–384.
[5] H. Zhang, L. Zhang, X. Qui, H. Li, P. H. Torr, and P. Koniusz, "Few-shot action recognition with permutation-invariant attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 525–542.
[6] W. Zhu, J. Lu, J. Li, and J. Zhou, "DSNet: A flexible detect-to-summarize network for video summarization," *IEEE Trans. Image Process.*, vol. 30, pp. 948–962, 2021.
[7] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3278–3292, Aug. 2021.
[8] B. Zhao, X. Li, and X. Lu, "TTH-RNN: Tensor-train hierarchical recurrent neural network for video summarization," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3629–3637, Apr. 2021.
[9] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 540–555.
[10] R. Anirudh, A. Masroor, and P. Turaga, "Diversity promoting online sampling for streaming video summarization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3329–3333.
[11] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2714–2721.
[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[13] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 766–782.
[14] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 863–871.

[15] B. Zhao, X. Li, and X. Lu, "HSA-RNN: Hierarchical structure-adaptive RNN for video summarization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7405–7414.

[16] L. L. Casas and E. Koblents, "Video summarization with LSTM and deep attention models," in *Proc. Int. Conf. MultiMedia Modeling*, 2019, pp. 67–79.

[17] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 39–54.

[18] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, "Deep audio-visual learning: A survey," 2020, *arXiv:2001.04758*. [Online]. Available: http://arxiv.org/abs/2001.04758

[19] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 208–224.

[20] Y.-B. Lin and Y.-C. F. Wang, "Audiovisual transformer with instance attention for audio-visual event localization," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 274–290.

[21] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2201–2207.

[22] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–11.

[23] X. Li, B. Zhao, and X. Lu, "Key frame extraction in the summary space," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1923–1934, Jun. 2018.

[24] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3652–3664, Aug. 2017.

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

[26] Y. Cong, J. Liu, G. Sun, Q. You, Y. Li, and J. Luo, "Adaptive greedy dictionary selection for web media summarization," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 185–195, Jan. 2017.

[27] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. Int. Conf. Image Process.*, 1988, pp. 866–870.

[28] S. E. F. de Avila, A. P. B. Lopes, A. da Luz, Jr., and A. de A. Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognit. Lett.*, vol. 32, no. 1, pp. 56–68, Jan. 2011.

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

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

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

[32] S. Mei, G. Guan, Z. Wang, M. He, X.-S. Hua, and D. D. Feng, "$L_{2,0}$ constrained sparse dictionary selection for video summarization," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.

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

[34] M. Ma, S. Mei, S. Wan, J. Hou, Z. Wang, and D. D. Feng, "Video summarization via block sparse dictionary selection," *Neurocomputing*, vol. 378, pp. 197–209, Feb. 2019.

[35] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[36] J. Ren, J. Jiang, and Y. Feng, "Activity-driven content adaptation for effective video summarization," *J. Vis. Commun. Image Represent.*, vol. 21, no. 8, pp. 930–938, Nov. 2010.

[37] N. Ejaz, T. B. Tariq, and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism," *J. Vis. Commun. Image Represent.*, vol. 23, no. 7, pp. 1031–1040, Oct. 2012.

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

[39] Y. J. Lee and K. Grauman, "Predicting important objects for egocentric video summarization," *Int. J. Comput. Vis.*, vol. 114, no. 1, pp. 38–55, 2015.

[40] Y. Jae Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1346–1353.

[41] S. Tschiatschek, R. Iyer, H. Wei, and J. Bilmes, "Learning mixtures of submodular functions for image collection summarization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1413–1421.

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

[43] J. Wang, W. Wang, Z. Wang, L. Wang, D. Feng, and T. Tan, "Stacked memory network for video summarization," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 836–844.

[44] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 347–363.

[45] J. Gao, X. Yang, Y. Zhang, and C. Xu, "Unsupervised video summarization via relation-aware assignment learning," *IEEE Trans. Multimedia*, vol. 23, pp. 3203–3214, 2021.

[46] J. Wang *et al.*, "Query twice: Dual mixture attention meta learning for video summarization," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4023–4031.

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

[48] K. Zhang, K. Grauman, and F. Sha, "Retrospective encoders for video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 383–399.

[49] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7582–7589.

[50] B. Pan *et al.*, "Spatio-temporal graph for video captioning with knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, p. 10.

[51] M. Dzabraev, M. Kalashnikov, S. Komkov, and A. Petiushko, "MDMMT: Multidomain multimodal transformer for video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3354–3363.

[52] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 164–172.

[53] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 3626–3637, 2020.

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

[55] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[56] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.

[57] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7596–7604.

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

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

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

[61] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1059–1067.

[62] H. Wei, B. Ni, Y. Yan, H. Yu, X. Yang, and C. Yao, "Video summarization via semantic attended networks," in *Proc. 32th AAAI Conf. Artif. Intell.*, 2018, pp. 216–223.

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

[64] J. Lei, Q. Luan, X. Song, X. Liu, D. Tao, and M. Song, "Action parsing-driven video summarization based on reinforcement learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 2126–2137, Jul. 2019.

[65] Y. Chen, L. Tao, X. Wang, and T. Yamasaki, "Weakly supervised video summarization by hierarchical reinforcement learning," in *Proc. ACM Multimedia Asia*, 2019, pp. 1–6.