DSTC-Sum: A Supervised Video Summarization Model Using Depthwise Separable Temporal Convolutional

M. Hamza Eissa[1], Hesham Farouk[2], Kamal Eldahshan[3], Amr Abozeid[4]
Department of Mathematics-Faculty of Science, Al-Azhar University, Cario, Egypt[1, 3, 4]
Department of Computers and Systems Electronics Research Institute, Cario, Egypt[2]
Department of Computer Science-College of Computer and Information Sciences, Jouf University, Saudi Arabia[3, 4]

***Abstract*—The exponential growth in video content has created a critical need for efficient video summarization techniques to enable faster and more accurate information retrieval. Video summarization has excellent potential to simplify the analysis of large video databases in various application areas ranging from surveillance, education, entertainment, and research. DSTC-Sum, a novel supervised video summarization model, is proposed based on Depthwise Separable Temporal Convolutional (DSTC). Leveraging the superior representational efficiency of DSTCN, the model addresses computational challenges and training inefficiencies encountered in traditional recurrent architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs). Additionally, this approach reduces computational overhead and memory usage. DSTC-Sum achieved state-of-the-art performance on two commonly used benchmark datasets, TVSum and SumMe, and outperformed all previous methods with F-scores by 1.8% and 3.33%, respectively. To validate the model's generality and robustness, the model was further tested on the YouTube and Open Video Project (OVP) datasets. The proposed model did slightly better on these datasets than several popular techniques, with F scores of 60.3 and 58.5, respectively. Finally, these findings confirm that this model captures long-term temporal dependencies and produces high-quality video summaries across all types of videos.**

***Keywords*—Video summarization; depthwise separable temporal convolutional; video processing; deep learning**

I. INTRODUCTION

In recent years, the proliferation of video capture devices and their declining costs have led to an unprecedented increase in video data volume. There are many kinds of visual data, but the video is one of the most significant. It is impossible to expect people to be able to see these videos and extract relevant information from them due to the vast amount of data included inside them. According to the Cisco Visual Networking Index report [1], it will take a human more than 5 million years to watch all the movies published on the Internet each month by 2022. Because of this, developing computer vision systems that effectively browse vast amounts of video data is becoming an increasingly important goal. Video summarizing has emerged as a potential technique that may assist viewers in dealing with the massive amount of data in the video.

When given an original video as input, video summarizing produces a more condensed version that still contains all the essential information from the original. Video summarizing has numerous potential applications (for example, indexing, browsing, and surveillance) [2, 3]. Summary videos may also be helpful for various downstream video analysis activities. For instance, running other analytic algorithms on shorter videos, such as action recognition, can be done more quickly.

Recent methodologies [4-6] address video summarization as a sequence labeling challenge, focusing on identifying and extracting key video segments efficiently. The Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) have been modified as an RNN variant to address this issue [7]. The LSTM model has a one-to-one correspondence between each time step and video frame. The LSTM model generates an output binary value at each time step, which indicates whether this frame was chosen for inclusion in the summary video. The LSTM methodology has the advantage of recording the long-range structural connections between frames. However, some limitations are embedded into these LSTM-based models. In LSTM, the computation typically proceeds from left to right. It indicates that the model can only perform one frame at a time, with each frame having to wait until the processing of the frame has been completed before it can begin. Even if there is bi-directional LSTM (Bi-LSTM) [8], the computation still has the same issue when using Bi-LSTM in either manner. Because of the sequential nature of the LSTM computation, it is impossible to readily parallelize it to make the most of the GPU resources. As temporal classifiers, sequence-based architectures such as RNN and LSTM are computationally costly, memory-heavy, and challenging to train. Temporal Convolutional Networks (TCNs) [9, 10] have recently demonstrated promising performance in video summarizing tasks.

In this work, to solve the abovementioned challenges, the DSTC-Sum model was designed based on the Depthwise Separable Temporal Convolutional Network (DSTCN), which efficiently extracts long-term temporal dependencies and local features. Unlike RNN-based models, which are computationally inefficient, sequential in nature, and reliant on domain-specific annotations, TCNs allow the addition of new layers while being computationally less expensive, quicker to train, and lightweight [11]. This makes traditional methods like RNNs and LSTMs unsuitable for large-scale datasets and

longer videos. The DSTC-Sum model leverages Depthwise Separable Convolutions (DSC) to improve feature representation while requiring low computation and memory [12]. Its scalable and dataset-agnostic design ensures efficient extraction of temporal dependencies and enhances generalizability across diverse video datasets. The benefits of DSTCN include the use of large kernels to capture long-range relationships while limiting the number of overall parameters, resulting in compact models. Additionally, large adjacent kernels effectively extract both global and local temporal features. By enabling simultaneous analysis of all video frames through GPU parallelization, DSTC-Sum addresses the challenges of computational inefficiency, scalability, and poor temporal modeling, achieving superior performance compared to current video summarization techniques.

We conducted comprehensive evaluations on two benchmark datasets, TVSum and SumMe. In the standard-supervised setting, the DSTC-Sum model achieved an F-score of 48.7%, which increased to 52.8% with data augmentation. On the TVSum dataset, the model attained an F-score of 61.2% in the standard setting, improving to 62.9% with augmentation. These results demonstrate the superior performance of the DSTC-Sum model compared to state-of-the-art techniques, with notable improvements of 1.8% and 3.33% on the two datasets, respectively. This highlights the model's enhanced ability to accurately predict the importance of video segments and generate high-quality summaries. The DSTC-Sum model's effective capture of temporal dependencies and key factors sets it apart, underscoring its potential for broader video summarization applications. To further evaluate the model's generalizability and robustness, we extended our experiments to two additional datasets: YouTube and the OVP. The model demonstrated superior performance on these datasets, achieving F-scores of 60.3% and 58.5%, respectively, outperforming several state-of-the-art techniques. These results underscore the model's effectiveness in capturing long-term temporal dependencies and generating high-quality summaries across various video genres.

The paper is organized as follows: Section II briefly discusses related studies on deep learning-based video summarizing approaches. Section III introduces our suggested DSTC-Sum model. Section IV discusses the model implementation and experimental findings. Lastly, the paper's conclusion is presented in Section V.

II. Related Work

Video summarization presents the challenge of selecting the most relevant segments of a video for inclusion in the summary and accurately identifying and extracting those segments from the entire video. This process requires a comprehensive understanding of the video content to ensure the summary represents the original video's essential aspects. In the early stage of video summarizing research, most approaches focus on a particular category of videos. For instance, the significance of a specific occurrence during a video segment of a show airing a sporting event can be easily determined by referring to the regulations governing that sport. [13]. In addition, certain sports games, such as baseball and American football, have a specific structure that makes extracting crucial segments of the game's action easier. Similarly, characters who feature in movies can also be domain knowledge [14]. In these areas, video summaries can be generated with the assistance of many kinds of metadata [15, 16]. Videos focusing on the creator alone are another fascinating example of video domains. A video summarization approach has been proposed using specific domain knowledge that can be considered a set of predetermined objects [17]. This approach aims to summarize videos in a manner that considers the domain's specifics. Newer methods in this general area use supervised learning techniques to incorporate domain knowledge. For instance, [18] offered to summarize a video with the primary focus on a particular event and use an event classifier's confidence score to measure a video segment's significance. However, due to the heavy reliance that such methods have on specific industry expertise, it is nearly impossible to generalize them to other types of writing.

When given an original video as input, the video summarizing goal is to produce a condensed version highlighting the most vital information from the original. There have been many other ways that this issue has been represented, such as in a video overview [19], time-lapses [20-22], montage [23, 24], and storyboards [25-29]. Our work is most closely associated with storyboards, consisting of a selection of a few typical frames of video that outline important events throughout an entire film. Storyboard-based summarization can produce two different kinds of outputs: keyframes [30], in which specific isolated frames are selected for forming the summary of the video, as well as key shots [31, 32], a method for generating a resume that considers a series of successive correlated frames contained within a temporal slot. Both types of outputs are referred to as keyframes.

Initial efforts in a video summarizing primarily rely on hand-crafted heuristics. Most of these methods do not require supervision. They specify a variety of heuristics to reflect the significance of the frames' representativeness [33-39], and they utilize the significance scores to select representative frames to form the video summary. Recent research has investigated supervised learning methodologies for video summaries [40-42]. These methods use video training data and the ground-truth summaries humans create for those videos. These supervised learning algorithms perform better than the early work on unsupervised methods because they can acquire sophisticated semantic knowledge that humans implicitly use to construct summaries.

Deep learning approaches have recently been popular for vision tasks, especially video summarization [43-45]. The foundation of LSTM is the theory that it can effectively capture long-range dependencies between video frames, which are necessary for creating insightful summaries. Zhang et al. [32] model the variable range dependency with two LSTMs and consider the video summarizing assignment of a problem of structured prediction based on data that can be sequential. Two Long Short-Term Memory (LSTM) networks are employed to analyze video sequences comprehensively. One LSTM is dedicated to processing the sequences in the forward direction, capturing the temporal dynamics as they unfold chronologically. Meanwhile, the other LSTM handles sequences in the reverse direction, allowing for a holistic

understanding of the video content from both temporal perspectives. They incorporate a determinantal point process model to improve further the subset selection's diversity [9, 46]. Mahasseni and colleagues present an unsupervised generative adversarial system consisting of the discriminator and summarizer [6]. The summarizer, an LSTM variational autoencoder, selects frames from the video and decodes the output to reconstruct the video. The discriminator is another LSTM network that gains the ability to distinguish between candidates by differentiating between the input video and its reconstruction. It accomplishes this by examining the variations between the two. They also incorporate a keyframe regularization into their algorithm, expanding it to supervise learning.

Despite significant advancements in video summarization, several gaps still need to be discovered in existing studies that necessitate further investigation. Prior approaches, such as those based on Recurrent Neural Networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), have proven effective in capturing temporal dependencies but are hindered by high computational costs, memory intensiveness, and sequential processing limitations, making them less scalable for larger datasets or longer video sequences. Additionally, evaluations in previous research are often confined to limited datasets, such as TVSum and SumMe, which do not adequately represent the diversity of real-world video content. The reliance on domain-specific annotations and handcrafted features further restricts the generalizability of these methods. While Temporal Convolutional Networks (TCNs) offer an alternative by addressing some of these limitations, there remains a need for lightweight, scalable architectures that combine computational efficiency with robust temporal modeling capabilities. This study addresses these gaps by proposing DSTC-Sum, a novel video summarization model based on Depthwise Separable Temporal Convolutions, which enhances efficiency and scalability, demonstrates generalizability across diverse datasets, and outperforms state-of-the-art methods in terms of F-scores and computational performance, thereby contributing to the advancement of efficient and robust video summarization techniques.

III. The Proposed Approach

This section introduces the DSTC-Sum model to summarize the input videos. Fig. 1 depicts the structured steps of the DSTC-Sum. First, the feature descriptors are generated from the input video frames using VGG16 [47]. Then, these feature vectors are fed into a series of Depthwise Separable Temporal Convolutional Blocks (DSTCB) that predict a score for each frame. Suppose that X represents a feature vector as $X_{1:n} = \{f_1, f_2, f_3, \ldots, f_n\}$. The model goal is to assign a corresponding score for each frame $Y_{1:n} = \{y_1, y_2, y_3, \ldots, y_n\}$, where n is the frame number, which varies depending on the video.

In the following subsections, we will describe the baseline model VGG16 and then explain the DSTCB, which is built using residual depthwise dilated blocks.

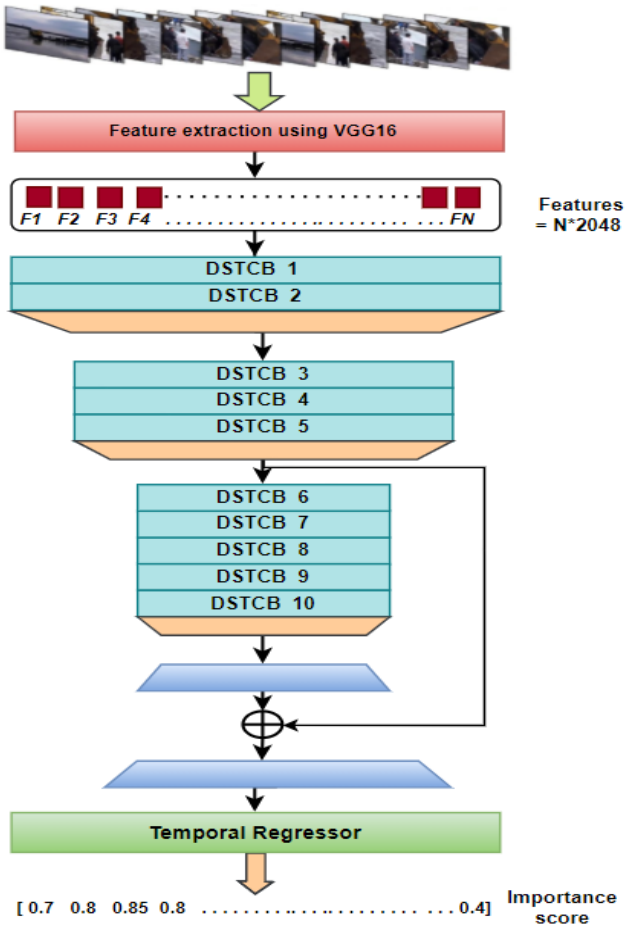


Fig. 1. The detailed structure of the DSTC-Sum.

A. Feature Extraction using VGG16

We start by feeding an input video into a feature extractor. The feature extractor module comprises the pre-trained first ten convolution layers of the VGG16 [47]. Because of its high generalization capabilities, the VGG16 is commonly employed as a feature extractor for many deep-learning models. We applied VGG16 to extract the basic features for the suggested model.

B. Depthwise Separable Temporal Convolutional Blocks (DSTCB)

Temporal convolutional network (TCN) is a CNN variant utilized in sequencing-based tasks and has recently outperformed alternative recurrent models like LSTM and gated recurrent units (GRUs) [48]. TCN enables temporal solid information extraction from sequential data [49]. TCNs aim to encapsulate temporal relationships with a broader receptive field. To capture large receptive fields, either a) dilations on consecutive TCN layers or b) big standard neighboring kernels are used. Dilated convolutions on consecutive layers help capture a broad temporal representation while reducing the number of training parameters. However, when additional layers are added, the kernels become increasingly sparse, resulting in the gridding artifacts issue.

When the dilation parameter increases at higher levels, the input data sample becomes increasingly sparse (Gridding Artifacts issue). As a result, local dependencies between neighboring pixels are lost, and the output layer is not temporally associated with their input sample. DS-TCN [10] proposes techniques for methodical aggregation of convolution layers in the following layers with constant or configurable dilation rates to address the issue of gridding artifacts. Therefore, each DSTCB consists of stacked 'N' depthwise dilated 1d temporal convolution layers.

Fig. 2 depicts the detailed construction of the DSTCB. All output levels receive a combined input from the preceding layer with various dilation rates. Each output layer is combined along a channel dimension to produce the NxC dimension. The cross-channel correlation is then estimated using a pointwise convolution procedure that decreases the dimensions of the channel from N×C to C from the concatenated data. This output is normalized using layer normalization [50]. To maintain adequate gradient flow, we also employ residual connections. Finally, we employ ReLU.

The DSTCB has various advantages:

1) Because depthwise convolutions are computationally efficient, we may employ huge-size kernels. As a result, we may employ lower dilations in conjunction with lengthy kernels to capture lengthy temporal features.

2) Scale information is recorded in each layer with varied dilation rates. Multi-scale information is contained in concatenated pointwise convolution. As a result, the model can tolerate different temporal durations for each event.

3) Stacking the outputs of all layers aids in data smoothing and eliminates the artifact effect caused by gridding. This method enables the model to acquire more detailed local features.

4) The receptive field can temporally extend without boosting the parameters by adjusting the block's dilation rates.

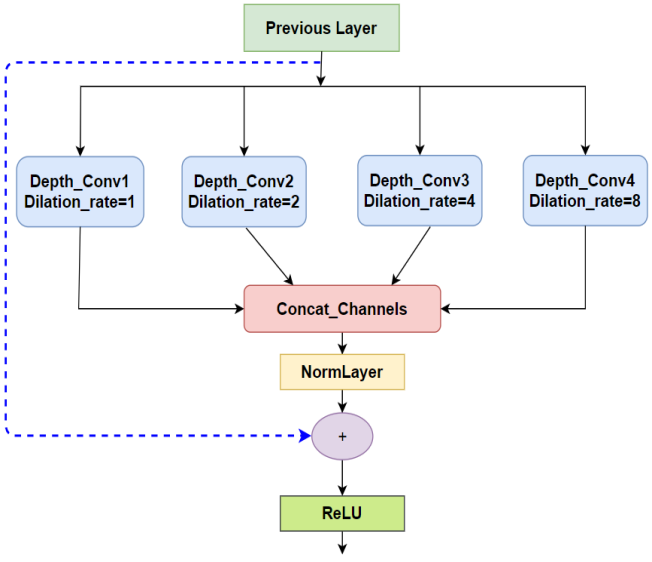


Fig. 2. The detailed construction of the DSTCB.

A temporal max-pooling was added after each DSTCB. Next, we take the output of DSTCB5, perform a 1x1 convolution layer and batch normalization, and then combine it with the production of deconv1 by element-wise addition. This merger is equivalent to the skip-connection in study [9]. Skip connections mix coarse and fine feature maps in semantic segmentation to acquire richer visual characteristics. This skip connector will also be valuable in video summarization, as it will aid in the recapture of temporal information needed for summary. Then, we do another temporal deconvolution to obtain the final representation of length N. The generated representation is fed into the temporal regressor network as input. Finally, the Regressor generates a collection of frame-level scores that represent the importance of the frames.

IV. EXPERIMENTS

This section provides an overview of the datasets, implementation details, and evaluation metrics used to assess the DSTC-Sum model. It outlines the training setup, key parameters, and performance criteria. The section concludes with a presentation and discussion of the DSTC-Sum results, highlighting its strengths, limitations, and potential areas for improvement.

A. Evaluation Datasets

The benchmark datasets for testing and evaluating our DSTC-Sum model are SumMe [51] and TVSum [52]. Table I shows several features of the target datasets. The SumMe benchmark dataset consists of 25 videos in a video benchmark dataset that has numerous topics and occurrences (like holidays, Sports, etc.). The videos on SumMe can vary between 1.5 to 6.5 minutes. The TVSum dataset consists of 50 videos from the TRECVid Multimedia Event Detection (MED) challenge [53]. The videos are divided into different categories, e.g., 'saucing up a sandwich,' 'showing dog,' and so on. This dataset houses videos that range from one minute to five minutes.

Previous research in [32] suggests that more videos should be added to the datasets to minimize the difficulty of training a deep neural network with a few manually annotated examples. Therefore, we bolster the existing training data by incorporating 39 videos from the YouTube dataset [54] in addition to 50 videos from the OVP dataset [54, 55]. The YouTube dataset includes some videos, including cartoons, sports, news, etc. OVP has videos in many different genres, such as documentaries. Because the multiple datasets provide ground-truth annotations in various formats, we adapt a training process that uses summaries based on keyframes to construct a unified set of ground truths for each video included in the datasets, as in study [32, 56].

TABLE I. OVERVIEW OF THE TARGET DATASETS

Datasets	Videos	Annotations	Duration (Min)
TVSum	50	20	2-10
SumMe	25	15-18	1-6
YouTube	39	5	1-10
OVP	50	5	1-4

B. Implementation and Training Details

The model is implemented using PyTorch, with the number of DSTCB blocks set to ten. Each DSTCB block consists of four depthwise one-dimensional temporal convolution layers, where the dilation parameter is doubled in each layer. We adopt the hyperparameters used in MS-TCN [12] to ensure a fair comparison with other methods. The Adam optimizer is employed with a learning rate of 0.0005. The parameters for the smoothing loss function are set as follows: smoothing λ=0.15 and threshold τ=4.

We downsample the videos uniformly to 2 frames per second for feature extraction, as done in [2]. We select representative frames from each video to reduce the final feature dimension to 320. Training frames are scaled to maintain consistent spatial dimensions across all videos. The DSTC-Sum model can handle longer videos and videos of varying lengths. We use the output from the maxpool5 layer of a pre-trained VGG16 model [47] as the feature descriptor for each frame, with a feature dimension of 512. Notably, our model is flexible and can work with any feature representation.

During training, we set the batch size to 5, the learning rate to 10^{-3}, and the momentum to 0.9. The Stochastic Gradient Descent (SGD) optimizer was used to train the DSTC-Sum model.

C. Evaluation Metrics

We evaluate the DSTC-Sum use of a keyshot-based metric, as in [6, 32]. Suppose that S_G is the ground-truth summary and S_E is the extracted summary for video V. We define the precision (P) and recall (R) by utilizing the temporal overlap between them as in Eq. (1) and (2):

$$P = \frac{|S_E \cap S_G|}{|S_E|} \quad (1)$$

$$R = \frac{|S_E \cap S_G|}{|S_G|} \quad (2)$$

As a final step, the evaluation is carried out utilizing the F-score, which is calculated in Eq. (3):

$$F = \frac{2P \times R}{P + R} \times 100 \quad (3)$$

D. Performance Analysis and Discussion

The performance of various summarization techniques on the SumMe dataset is outlined in Table II and visualized in Fig. 3. The proposed DSTC-Sum method significantly outperforms other state-of-the-art techniques across almost all parameters. Specifically, in the standard-supervised setting, DSTC-Sum achieves an F-score of 48.7, higher than the following best technique, SUM-FCN, which scores 47.5. DSTC-Sum shows an even more substantial improvement in the augmented setting, achieving an F-score of 52.8 compared to SUM-FCN's 51.1. This consistent outperformance highlights the effectiveness of DSTC-Sum in summarizing videos within the SumMe dataset.

Similarly, the results on the TVSum dataset, presented in Table III and Fig. 4, demonstrate the superior performance of the DSTC-Sum approach. DSTC-Sum achieves an F-score of 61.2 in the standard-supervised setting, edging out with close competitors like M-AVS and DHAVS, which scored 61.0 and 60.8, respectively. The augmented setting further showcases the dominance of DSTC-Sum, with an F-score of 62.9, significantly higher than M-AVS's 61.8 and SUM-GAN*sup*'s 61.2. These results underline the robustness and efficiency of DSTC-Sum in producing high-quality video summaries on the TVSum dataset.

TABLE II. SUMMARIZATION PERFORMANCE (F-SCORE) COMPARISON ON THE SUMME BENCHMARK DATASET BETWEEN DSTC-SUM AND OTHER TECHNIQUES USING DIFFERENT PARAMETERS

Technique	Standard-Supervised	Augmented
DPP-LSTM [5]	38.6	42.9
SUM-GAN*sup* [6]	41.7	43.6
Li et al. [57]	43.1	–
M-AVS [58]	44.4	46.1
DHAVS [59]	45.6	46.5
SUM-FCN [9]	47.5	51.1
DSTC-Sum	**48.7**	**52.8**

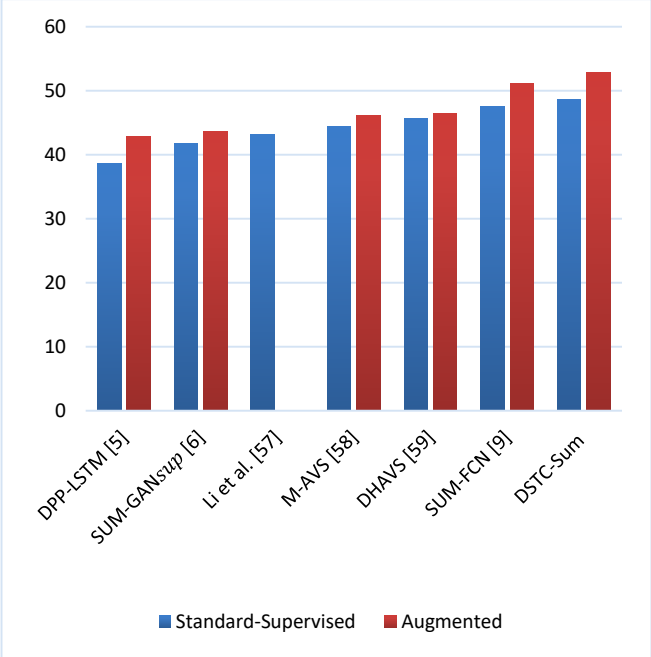


Fig. 3. Summarization performance (F-score) comparison on the SumMe dataset.

TABLE III. SUMMARIZATION PERFORMANCE (F-SCORE) COMPARISON ON THE TVSUM DATASET BETWEEN DEPTHTEMPORAL-SUM AND OTHER TECHNIQUES USING DIFFERENT PARAMETERS

Technique	Standard-Supervised	Augmented
DPP-LSTM [5]	54.7	59.6
SUM-GAN*sup* [6]	56.3	61.2
Li et al. [57]	52.7	–
SUM-FCN [9]	56.8	59.2
M-AVS [58]	61.0	61.8
DHAVS [59]	60.8	61.2
DSTCN-Sum	**61.2**	**62.9**

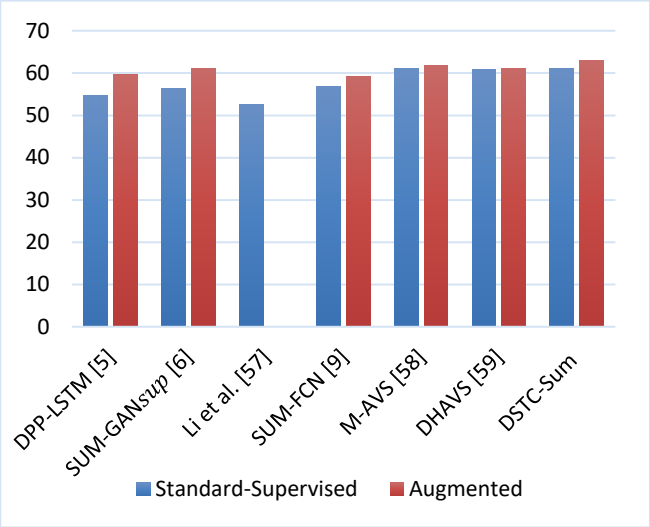


Fig. 4. Summarization performance (F-score) comparison on the SumMe dataset.

The DSTC-Sum methodology consistently demonstrates superior performance in video summarization tasks compared to existing state-of-the-art techniques. Its notable effectiveness, especially in augmented settings, indicates its high efficiency in improving summarization quality, as evidenced by the F-score metrics. The robust applicability of DSTC-Sum across different datasets further suggests its potential for wide adoption in video summarization applications. The implications of these findings are significant for the field of video summarization. By providing a method that consistently outperforms existing techniques, DSTC-Sum can enhance various applications, from creating more engaging video highlights for entertainment to improving the efficiency of video data management in professional and educational contexts. Moreover, the scalability and adaptability of DSTC-Sum means it could be integrated into various platforms and devices, including mobile applications and cloud-based services, thereby broadening its impact.

E. Extended Experiments

To further evaluate the effectiveness and generalizability of the DSTC-Sum model, we extended our experiments to include two additional benchmark datasets: YouTube and OVP (Open Video Project). These datasets were chosen due to their diverse content types, which pose unique challenges to video summarization models. By expanding our experimental scope, we aim to demonstrate the robustness of the proposed model across a broader range of video genres and characteristics. The DSTC-Sum model was fine-tuned on both datasets using the same configuration as in prior experiments. Specifically, the model architecture consisted of 10 Depthwise Separable Temporal Convolutional Blocks (DSTCB), with 4 depthwise convolutional layers per block. The Adam optimizer with a learning rate of 0.0005 was utilized, and the number of training epochs was adjusted based on the dataset size to prevent overfitting.

We applied data augmentation techniques such as random cropping and video flipping during training to improve generalization. Like the TVSum and SumMe datasets, the extracted frame-level features were fed into the model for training and evaluation. We benchmarked the model against several state-of-the-art video summarization models, including SUM-GAN, MS-TCN, and DPP-LSTM.

Table IV summarizes the YouTube dataset results. The proposed DSTC-Sum model achieved an F-score of 60.3%, outperforming the following best method, MS-TCN, which achieved an F-score of 58.6%. This improvement can be attributed to the model's ability to capture both long-term and short-term temporal dependencies, which is essential for summarizing the diverse content found in YouTube videos.

TABLE IV. PERFORMANCE COMPARISON OF VIDEO SUMMARIZATION METHODS ON THE YOUTUBE DATASET, MEASURED USING F-SCORE, PRECISION, AND RECALL

Technique	F-score (%)	Precision	Recall
DPP-LSTM [5]	55.2	54.8	55.7
SUM-GAN*sup* [6]	56.3	55.9	56.8
SUM-FCN [9]	58.6	57.9	59.2
DSTC-Sum	**60.3**	**60.1**	**60.6**

TABLE V. PERFORMANCE COMPARISON OF VIDEO SUMMARIZATION METHODS ON THE OVP DATASET, MEASURED USING F-SCORE, PRECISION, AND RECALL

Technique	F-score (%)	Precision	Recall
DPP-LSTM [5]	52.7	51.9	53.5
SUM-GAN*sup* [6]	56.3	55.4	57.0
SUM-FCN [9]	56.1	55.4	56.8
DSTC-Sum	**58.5**	**58.0**	**59.0**

Table V presents the results of the OVP dataset. Here, DSTC-Sum achieved an F-score of 58.5%, again outperforming the compared methods. With its more structured content, the OVP dataset benefited from the model's ability to capture long-range dependencies without losing important local features, a challenge that other models, such as SUM-GANsup, struggled with.

As shown in Fig. 5, the experimental results demonstrate the effectiveness of the DSTC-Sum model across both YouTube and OVP datasets. The model's ability to summarize videos of varying lengths and content types significantly influenced its performance in the YouTube dataset. The diversity in YouTube videos requires a model capable of understanding both global and local temporal structures, which is one of the key strengths of the DSTC-Sum model. On the OVP dataset, the model's performance highlights its ability to handle shorter, more structured videos. Compared to the baseline models, the improved F-score on this dataset shows that DSTC-Sum is particularly effective at summarizing videos with well-defined narrative structures, such as documentaries and educational videos.

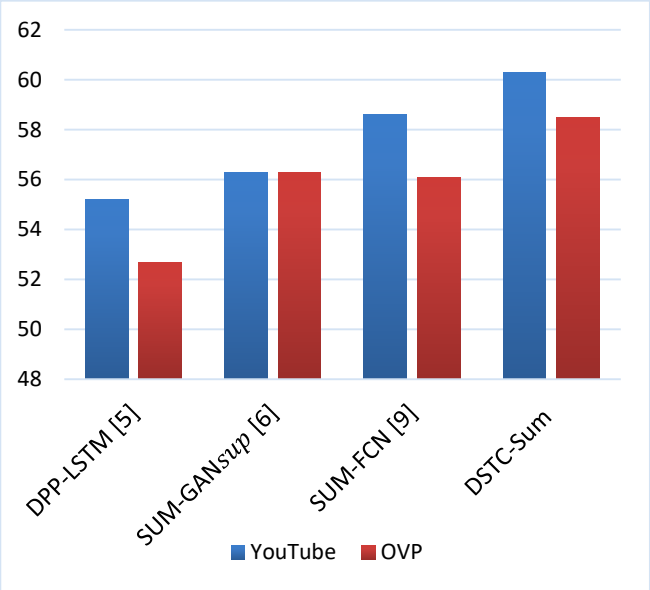


Fig. 5. DSTC-Sum model across both YouTube and OVP datasets.

The results on both YouTube and OVP datasets reinforce the generalizability and effectiveness of the DSTC-Sum model in video summarization tasks. The model consistently outperforms existing methods, demonstrating its ability to capture both long-term and short-term temporal dependencies. This makes it suitable for videos with diverse content (YouTube) and structured narratives (OVP).

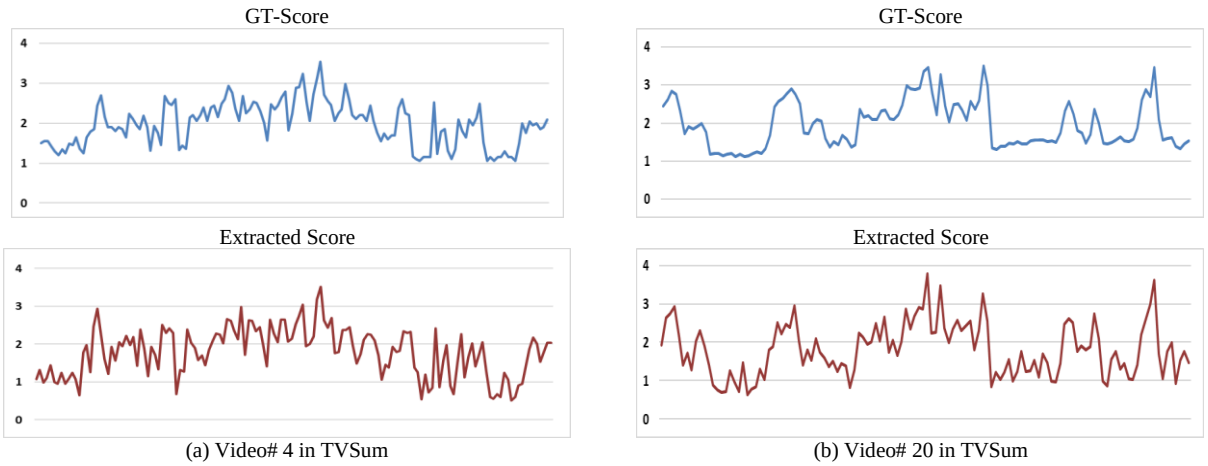
The comparison across all datasets (TVSum, SumMe, YouTube, and OVP) indicates that the DSTC-Sum model is robust and versatile in different summarization tasks, whether the videos are user-generated content (YouTube), educational (OVP) or professionally curated datasets (TVSum, SumMe). The scalability and low computational cost of DSTCB architecture further emphasize its potential for practical applications, including real-time video summarization.

F. Qualitative Results

In addition to the quantitative evaluations presented in previous sections, assessing the DSTC-Sum model's performance from a qualitative perspective is essential. This section provides a deeper insight into how effectively the model captures important segments and generates accurate video summaries. We aim to demonstrate the model's ability to identify critical video moments by visualizing the extracted importance and ground truth scores. As seen in Fig. 6, we plot the extracted importance scores and the ground truth scores for two videos from the TVSum dataset to understand better how well our DSTC-Sum has learned. The important ratings derived from ground truth and the extracted scores generated by the suggested DSTC-Sum roughly match. Moreover, the proposed technique produces high-quality video summaries as users incorporate several factors that our DSTC-Sum considers relevant.

To further understand how effective our DSTC-Sum is, we showcase some video summaries that demonstrate its temporal modeling capabilities compared to SUM-GAN*sup* and DHAVS [57]. Fig. 7 illustrates that colorful bars indicate projected video summaries and sky-blue bars indicate ground truth scores. The segments selected as summaries by SUM-GAN*sup*, DHAVS, and DSTC-Sum are shown by the yellow, green, and red bars, respectively.

The DSTC-Sum technique produces high-quality video summaries by capturing temporal dependencies, allowing it to identify the most crucial video segments effectively. Comparisons of the summaries generated by DSTC-Sum with those produced by SUM-GANsup and DHAVS reveal that while other approaches often fail to select the most relevant sub-shots, DSTC-Sum consistently identifies key segments with higher ground-truth relevance scores. This ability to create accurate and relevant video summaries underscores DSTC-Sum's superiority in video summarization tasks.



(a) Video# 4 in TVSum (b) Video# 20 in TVSum

Fig. 6. Video scores extracted by the proposed DSTC-Sum (bottom) and Ground truth scores (top) for video 4 and video 20 in the TVSum dataset.

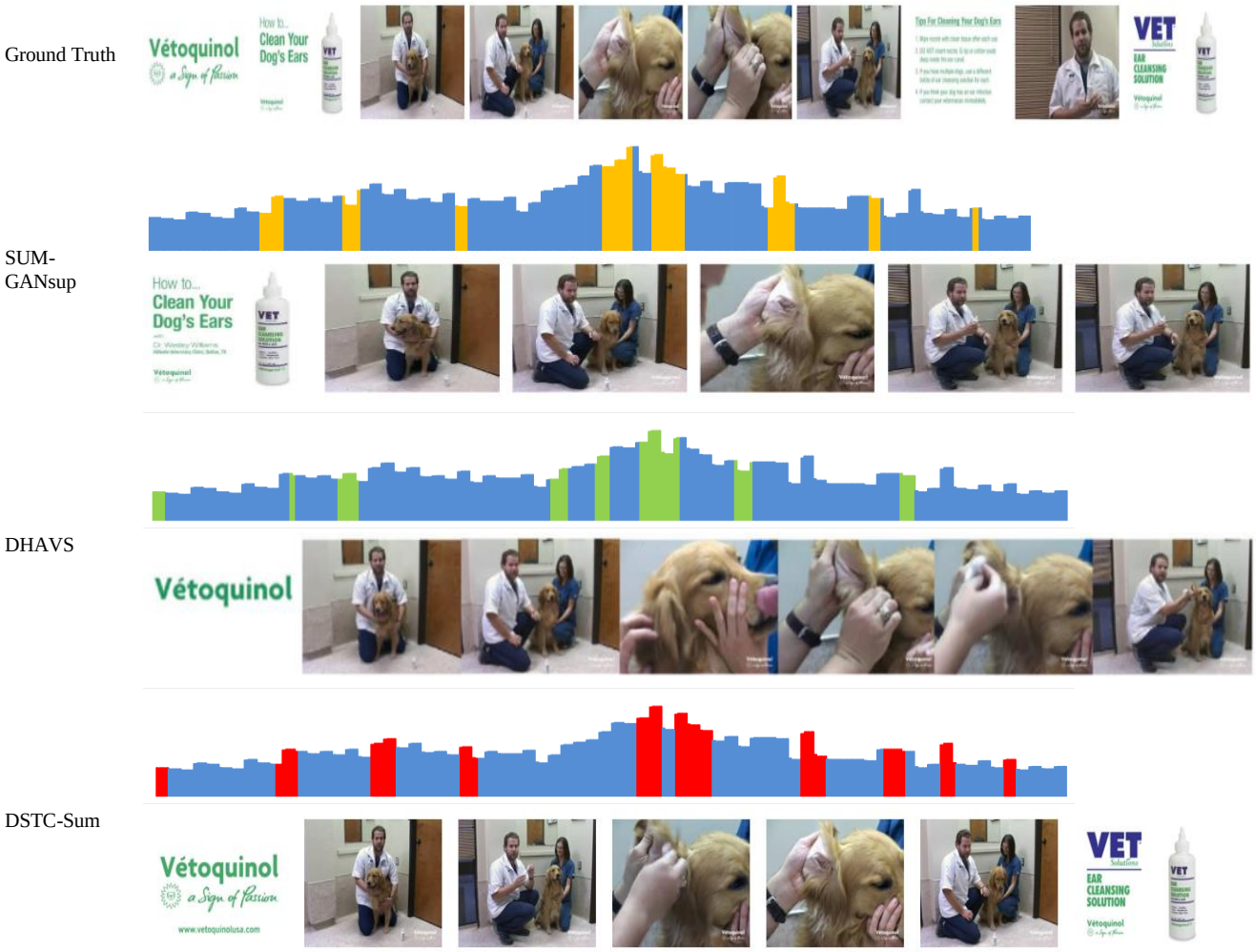


Fig. 7. Comparison of the extracted summary extraction for video 4 and video 20 in the TVSum dataset.

V. Conclusion

This study introduces DSTC-Sum, a video summarization model leveraging Depthwise Separable Temporal Convolutions (DSTC) to capture long-term temporal relationships and local features effectively. TCNs, unlike RNN-based models, allow for adding more layers while being computationally less expensive, faster to train, and lightweight. Furthermore, DSTC improves representational efficiency while being low-cost in computing and memory. Extensive experiments are carried out on two benchmark datasets, TVSum and SumMe. The outcomes demonstrate the efficacy of our DSTC-Sum model for supervised video summarization. Furthermore, the qualitative findings show that our model can produce fine-grained summary predictions and better scoring for each frame. To further assess the model's generalizability and robustness, we extended our experiments to two additional datasets: YouTube and the OVP (Open Video Project). On both datasets, the proposed model demonstrated superior performance, achieving F-scores of 60.3% and 58.5%, respectively, surpassing several state-of-the-art techniques. These results highlight the model's effectiveness in capturing long-term temporal dependencies and generating high-quality video summaries across various genres.

In the future, we plan to enhance the DSTC-Sum framework by integrating attention mechanisms to gain richer contextual information and improve summarization accuracy. Additionally, we will explore its potential in real-time video processing and personalized content creation, aiming to extend its application scope and solidify its position as a leading methodology in video summarization.

References

[1] X. Dong, Y. Yu, and J. Zhou, Cisco: Integration of Innovation and Operation. Springer Nature, 2023.

[2] E. Mofreh, A. Abozeid, H. Farouk, and K. A. ElDahshan, "Multi-object semantic video detection and indexing using a 3D deep learning model," Int. J. Intell. Eng. Syst., vol. 15, no. 3, 2022.

[3] K. A. ElDahshan, H. Farouk, A. Abozeid, and M. H. Eissa, "Global dominant SIFT for video indexing and retrieval," J. Theor. Appl. Inf. Technol., vol. 97, no. 19, pp. 5023–5035, 2019.

[4] H. Farouk, K. A. ElDahshan, and A. Abozeid, "Effective and efficient video summarization approach for mobile devices," Int. J. Interact. Mob. Technol., vol. 10, no. 1, 2016.

[5] J. Zhang, G. Wu, X. Bi, and Y. Cui, "Video summarization generation network based on dynamic graph contrastive learning and feature fusion," Electronics, vol. 13, no. 11, p. 2039, 2024.

[6] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[8] M. Sundermeyer, T. Alkhouli, J. Wuebker, and H. Ney, "Translation modeling with bidirectional recurrent neural networks," in Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP), Oct. 2014, pp. 14–25.

[9] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018.

[10] B. Hampiholi, C. Jarvers, W. Mader, and H. Neumann, "Depthwise separable temporal convolutional network for action segmentation," in Proc. Int. Conf. 3D Vis. (3DV), Nov. 2020, pp. 633–641.

[11] D. Li, R. Wang, P. Chen, C. Xie, Q. Zhou, and X. Jia, "Visual feature learning on video object and human action detection: A systematic review," Micromachines, vol. 13, no. 1, p. 72, 2021.

[12] G. Li, J. Zhang, M. Zhang, R. Wu, X. Cao, and W. Liu, "Efficient depthwise separable convolution accelerator for classification and UAV object detection," Neurocomputing, vol. 490, pp. 1–16, 2022.

[13] A. A. Khan and J. Shao, "SPNet: A deep network for broadcast sports video highlight generation," Comput. Electr. Eng., vol. 99, p. 107779, 2022.

[14] D. Zhao, D. Zhu, X. Min, J. Yue, K. Zhang, Q. Zhou, J. Zhao, and X. Yang, "Human attention-based movie summarization: Dataset and baseline model," Neurocomputing, vol. 534, pp. 106–118, 2023.

[15] P. Narwal, N. Duhan, and K. K. Bhatia, "A comprehensive survey and mathematical insights towards video summarization," J. Vis. Commun. Image Represent., vol. 89, p. 103670, 2022.

[16] T. Psallidas, M. D. Vasilakakis, E. Spyrou, and D. K. Iakovidis, "Multimodal video summarization based on fuzzy similarity features," in Proc. IEEE 14th Image, Video, Multidimensional Signal Process. Workshop (IVMSP), Jun. 2022, pp. 1–5.

[17] M. Dehghani, A. Gritsenko, A. Arnab, M. Minderer, and Y. Tay, "Scenic: A jax library for computer vision research and beyond," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 21393–21398.

[18] L. Fei-Fei and R. Krishna, "Searching for computer vision north stars," Daedalus, vol. 151, no. 2, pp. 85–99, 2022.

[19] R. Akhare and S. Shinde, "Query-focused video summarization: A review," in Artif. Intell. First Int. Symp. ISAI 2022, Haldia, India, Feb. 17–22, 2022, Revised Selected Papers, Springer, 2023.

[20] A. Markwirth, M. Lachetta, V. Mönkemöller, R. Heintzmann, W. Hübner, T. Huser, and M. Müller, "Video-rate multi-color structured illumination microscopy with simultaneous real-time reconstruction," Nat. Commun., vol. 10, no. 1, p. 4315, 2019.

[21] D. de Matos, W. Ramos, L. Romanhol, and E. R. Nascimento, "Musical hyperlapse: A multimodal approach to accelerate first-person videos," in *Proc. 34th SIBGRAPI Conf. Graph., Patterns, Images*, Oct. 2021, pp. 184–191.

[22] M. Silva, W. Ramos, A. Neves, E. Araujo, M. Campos, and E. R. Nascimento, "Fast-forward methods for egocentric videos: A review," in *Proc. 32nd SIBGRAPI Conf. Graph., Patterns, Images Tutorials (SIBGRAPI-T)*, Oct. 2019, pp. 36–46.

[23] F. Tian, J. Fan, X. Yu, S. Du, M. Song, and Y. Zhao, "TCVM: Temporal contrasting video montage framework for self-supervised video representation learning," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1539–1555.

[24] 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, 2020.

[25] J. A. Ghauri, S. Hakimov, and R. Ewerth, "Supervised video summarization via multiple feature sets with parallel attention," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2021.

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

[27] V. Kaushal, S. Subramanian, S. Kothawade, R. Iyer, and G. Ramakrishnan, "A framework towards domain-specific video summarization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 666–675.

[28] P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, "Exploring global diverse attention via pairwise temporal relation for video summarization," *Pattern Recognit.*, vol. 111, p. 107677, 2021.

[29] 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, 2020.

[30] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017.

[31] 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.*, 2016, pp. 1059–1067.

[32] K. Zhang, W. L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 11–14, 2016, Part VII, Springer Int. Publ., pp. 766–782.

[33] M. Fei, W. Jiang, and W. Mao, "Learning user interest with improved triplet deep ranking and web-image priors for topic-related video summarization," *Expert Syst. Appl.*, vol. 166, p. 114036, 2021.

[34] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T. S. Chua, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1437–1445.

[35] S. Lee, J. Sung, Y. Yu, and G. Kim, "A memory network approach for story-based temporal summarization of 360 videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1410–1419.

[36] J. Park, J. Lee, I. J. Kim, and K. Sohn, "Sumgraph: Video summarization via recursive graph modeling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, UK, Aug. 23–28, 2020, Part XXV, Springer Int. Publ., pp. 647–663.

[37] Y. Yuan, T. Mei, P. Cui, and W. Zhu, "Video summarization by learning deep side semantic embedding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 226–237, 2017.

[38] R. Panda and A. K. Roy-Chowdhury, "Collaborative summarization of topic-related videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[39] S. S. Zang, H. Yu, Y. Song, and R. Zeng, "Unsupervised video summarization using deep non-local video summarization networks," *Neurocomputing*, vol. 519, pp. 26–35, 2023.

[40] K. Zhang, K. Grauman, and F. Sha, "Retrospective encoders for video summarization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[41] A. Sharghi, A. Borji, C. Li, T. Yang, and B. Gong, "Improving sequential determinantal point processes for supervised video summarization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 517–533.

[42] M. Fei, W. Jiang, and W. Mao, "Creating memorable video summaries that satisfy the user's intention for taking the videos," *Neurocomputing*, vol. 275, pp. 1911–1920, 2018.

[43] 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, 2020.

[44] T. Liu, Y. Yuan, G. Teng, and X. Meng, "Improved deep convolutional neural network-based method for detecting winter jujube fruit in orchards," *Eng. Lett.*, vol. 32, no. 3, 2024.

[45] Y. Fu, L. Qiu, X. Kong, and H. Xu, "Deep learning-based online surface defect detection method for door trim panel," *Eng. Lett.*, vol. 32, no. 5, 2024.

[46] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[48] N. Lu, T. Yin, and X. Jing, "Deep learning solutions for motor imagery classification: A comparison study," in *Proc. 8th Int. Winter Conf. Brain-Comput. Interface (BCI)*, 2020.

[49] Y. K. Musallam, N. I. AlFassam, G. Muhammad, S. U. Amin, M. Alsulaiman, W. Abdul, and M. Algabri, "Electroencephalography-based motor imagery classification using temporal convolutional network fusion," *Biomed. Signal Process. Control*, vol. 69, p. 102826, 2021.

[50] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[51] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, Sep. 6–12, 2014, Part VII, Springer Int. Publ., pp. 505–520.

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

[53] P. Over and W. Kraaij, "Evaluation campaigns and TRECVID," in *Proc. ACM SIGMM Int. Workshop Multimedia Inf. Retrieval*, 2006.

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

[55] G. Geisler and G. Marchionini, "The open video project: Research-oriented digital video repository," in *Proc. Fifth ACM Conf. Digital Libr.*, 2000.

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

[57] 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, 2017.

[58] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder–decoder networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1709–1717, 2019.

[59] J. Lin, S.-h. Zhong, and A. Fares, "Deep hierarchical LSTM networks with attention for video summarization," *Comput. Electr. Eng.*, vol. 97, p. 107618, 2022.