Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Full length article

A deep audio-visual model for efficient dynamic video summarization[☆]

Gamal El-Nagar[a], Ahmed El-Sawy[a], Metwally Rashad[a,b,*]

[a] Department of Computer Science, Faculty of Computers & Artificial Intelligence, Benha University, Benha, Egypt
[b] Department of Computer Engineering and Information, College of Engineering, Wadi Ad Dwaser, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia

ARTICLE INFO

Keywords:
Video Skimming
VGGish
SumMe
TVSum
Visualization of score curves

ABSTRACT

The adage "a picture is worth a thousand words" resonates in the digital video domain, suggesting that a video could be seen as a composition of millions of these words. Videos are composed of countless frames. Video summarization creates cohesive visual units in scenes by condensing shots from segments. Video summarization gains prominence by condensing lengthy videos while retaining crucial content. Despite effective techniques using keyframes or keyshots in video summarization, integrating audio components is imperative. This paper focuses on integrating deep learning techniques to generate dynamic summaries enriched with audio. To address that gap, an efficient model employs audio-visual features, enriching summarization for more robust and informative video summaries. The model selects keyshots based on their significance scores, safeguarding essential content. Assigning these scores to specific video shots is a pivotal yet demanding task for video summarization. The model's evaluation occurs on benchmark datasets, TVSum and SumMe. Experimental outcomes reveal its efficacy, showcasing considerable performance enhancements. On the TVSum, SumMe datasets, an F-Score metric of 79.33% and 66.78%, respectively, is achieved, surpassing previous state-of-the-art techniques.

1. Introduction

In recent years, the exponential growth of digital video content has led to an increasing demand for efficient methods to process and summarize large-scale video datasets. Video summarization, as a subfield of multimedia analysis, aims to extract the most informative and important segments from a lengthy video, condensing its content while preserving its essential message. Such summarization is invaluable for facilitating efficient content browsing, storage, and retrieval and enhancing the user experience. Video summarization can be broadly categorized into two main types: static and dynamic summarization. Static summarization involves selecting a fixed number of keyframes or shots to represent the entire video [1–6]. However, this approach may lead to information loss and overlook crucial moments that occur between keyframes. On the other hand, dynamic summarization offers a more comprehensive and fluid representation by selecting key moments across the video timeline, providing a smoother playback experience [7–10]. Dynamic video summarization, while a powerful approach, is not without its limitations. The limitations of dynamic summarization become apparent when considering the need for coherent and informative representations of videos. Dynamic summarization faces challenges in maintaining temporal coherence, handling diverse content types, and avoiding information loss during the summarization process.

To achieve this objective, researchers have turned to deep learning, which has emerged as a powerful paradigm for solving complex tasks in various domains, including computer vision and natural language processing. The application of deep learning to video summarization has shown promising results, as it enables the automatic extraction of high-level features from visual and audio inputs. Video summarization techniques have witnessed remarkable advancements in leveraging deep learning capabilities. The proposed model's significance lies in successfully integrating audio and visual features, yielding comprehensive and contextually meaningful video summaries.

The proposed model in this paper involves a multi-phase process that leverages state-of-the-art neural networks and audio processing methods. Specifically, the video divided into shots and further extracts visual and audio features to represent each segment comprehensively. To capture the visual characteristics of the video, the powerful ResNet50 and InceptionV3 architectures are employed. These deep convolutional neural networks have shown exceptional performance in

[☆] This paper has been recommended for acceptance by Zicheng Liu.

[*] Corresponding author at: Department of Computer Engineering and Information, College of Engineering, Wadi Ad Dwaser, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia.

E-mail addresses: gamal.essam@fci.bu.edu.eg (G. El-Nagar), ahmed.el_sawy@fci.bu.edu.eg (A. El-Sawy), metwally.rashad@fci.bu.edu.eg (M. Rashad).

https://doi.org/10.1016/j.jvcir.2024.104130
Received 24 October 2023; Received in revised form 18 February 2024; Accepted 24 March 2024
Available online 28 March 2024
1047-3203/© 2024 Elsevier Inc. All rights reserved.

various image-related tasks and can efficiently extract visual features from individual video frames. Moreover, recognizing the significance of audio in video content, Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-Spectrogram are used for audio representation. Additionally, the proposed model incorporates the VGGish model, designed explicitly for audio representation, to capture the essential audio features within the video.

The integration of audio-visual features plays a pivotal role in the proposed model. Considering both modalities, the proposed model aims to create summaries that are more informative and engaging summaries than traditional video summarization methods, focusing solely on visual cues. The crux of the proposed model lies in employing an Artificial Neural Network (ANN) to predict the importance scores of individual shots based on the fused audio-visual features. This dynamic scoring process enables us to generate summaries that highlight crucial moments in the video while maintaining context and coherence. Our motivation stems from the pressing need to advance video summarization techniques beyond their current limitations. Traditional video summarization methods [4,11–13] which primarily focus on visual content. This lack of audio in multimedia experiences hinders holistic understanding and engagement. The research aims to bridge this gap by ensuring video summaries are not only visually informative but also rich with an audio narrative, enhancing the user experience and enhancing multimedia experiences. The main contributions of this paper are summarized as:

1. A proposed model for dynamic video summarization is presented to achieve a significant breakthrough. The model seamlessly integrates audio and visual features. It creates video summaries that are more comprehensive and contextually meaningful compared to traditional methods.
2. Leveraging state-of-the-art neural networks, including ResNet50, InceptionV3, and the specialized VGGish model, is crucial. This ensures a robust representation of visual and audio features. Consequently, it enhances the overall quality of the summarization process.
3. The proposed model's core innovation lies in the dynamic scoring process facilitated by an Artificial Neural Network (ANN). This enables the accurate prediction of importance scores for individual shots.
4. The performance of the proposed model is evaluated using SumMe and TVSum datasets. Standard performance measures, such as precision, recall, and f-score, are employed for this assessment. The results are then compared with state-of-the-art methods that use the same dataset source.

The rest of this paper follows this organization: Section 2 reviews related work in video summarization. Section 3 details the proposed dynamic audio-visual model, including shot segmentation, feature extraction, and the deep learning model for shot importance prediction. Section 4 presents experimental results and discusses the performance of the proposed model. Finally, Section 5 provides conclusions and outlines potential future research directions.

2. Related works

This section provides a comprehensive overview of the existing literature and research in the field, serving as a foundation for understanding the context and significance of video summarization. This paper delves into prior work's key findings, methodologies, and contributions, highlighting the gaps and opportunities that proposed model aims to address. [6] devises a method for generating static video summaries of input videos. The scope of this paper is meticulously tailored to accommodate users' preferences for concise yet informatively rich summaries. The effectiveness of the proposed technique is demonstrated seamlessly across an array of video genres, consistently ensuring the delivery of reliable and consistent outcomes. In pursuing this objective, the frame set corresponding to the input video is subjected to a series of processing steps meticulously designed to eliminate redundant frames. High-level feature vectors are extracted from this carefully optimized frame set, with the instrumental involvement of Convolutional Neural Networks (CNN) in their utilization. The effective representation of frames is achieved through the astute leveraging of the activations within the fully connected layers of pre-trained CNN models. To further elevate the quality of these feature vectors, the extracted features from four distinct pre-trained CNN models are thoughtfully amalgamated, creating a harmonious and comprehensive representation. This amalgamation process culminates seamlessly, feeding these feature-rich representations into a sophisticated Sparse Autoencoder (SAE). The role of the SAE is to refine and enrich the feature vectors, ensuring the encapsulation of every critical detail with precision and finesse. At the ultimate stage of this meticulously orchestrated process, a discerning Random Forest classifier is called into action. Its mission is to meticulously identify the keyframes, serving as the vanguard for generating a static video summary that retains the essence of the content's most crucial aspects. In sum, the paper's method has been systematically designed to ensure that users are presented with succinct yet profoundly informative video summaries underpinned by a sophisticated technological framework.

The authors in [14] present the architecture of the "Multi-Source Visual Attention (MSVA)" model, combines content-based image features and motion-related features using attention mechanisms and fusion techniques to assign importance scores to frames for video summarization. The choice of fusion method impacts the model's performance in selecting keyframes. The model's primary goal is to assign importance scores to video frames. It uses different features, including content-based image features from GoogleNet, motion features from the I3D model (RGB and optical flow), and an attention mechanism. The attention module processes these features. An aperture window and attention weights are calculated for each frame based on its relationship with others in the sequence. After attention, the latent representations of these features are fused through addition. This fused vector is processed through several neural network layers, including linear transformations, normalization, activation functions, and dropout layers, culminating in the prediction of importance scores for each frame using a sigmoid function. The model explores different fusion techniques, including early and late fusion, to determine the most effective way to combine the feature types.

The paper's authors [15] present the AudioVisual Recurrent Network (AVRN), a system that seamlessly integrates audio and visual data into the video summarization process. The AVRN comprises three fundamental elements. Firstly, it employs a two-stream LSTM to encode audio and visual features sequentially, thereby capturing their temporal relationships. Secondly, an audiovisual fusion LSTM dynamically combines audio and visual data to promote consistency and reduce disparities. Lastly, a self-attention module is introduced to capture global video information. While the two-stream LSTM processes audio and visual data separately, encoding temporal dependencies within each data type, it does not adequately address the discrepancies between these modalities, which could potentially disrupt the video summarization task. To overcome this issue, the paper presents an audiovisual fusion LSTM. This component takes combined audio and visual data as input and employs an adaptive gating mechanism to regulate the flow of information between the two modalities. Additionally, a self-attention video encoder is implemented to handle the intricacies of video structure. It identifies global dependencies among frames, primarily when no clear temporal connections exist between consecutive shots. This module enhances the sequence-based model by considering the video's overall structure and content. Regarding summary generation, the AVRN calculates the significance of each frame by incorporating fused audiovisual data, temporal dependencies, and global video information. Shot-level importance is determined by averaging frame-level significance within each shot, and video summaries are generated based on these shot-level importance scores.

The authors in [9] outlines the architecture and methodology of a video summarization model based on VASNet. This model employs a soft self-attention mechanism to capture frame dependencies, producing importance estimates for each frame. A novel method is introduced to transition from supervised to unsupervised learning. It focuses on specific attention matrix parts corresponding to non-overlapping video fragments, estimates their significance, creates a block diagonal sparse attention matrix to reduce parameters, and uses a simple loss function related to the summary length. The pipeline includes deep feature extraction using a pretrained CNN model and attention mechanism calculations for Query, Key, and value matrices, followed by concentrated attention on non-overlapping blocks. It also computes attentive uniqueness and diversity values for each frame. The resulting features are combined with original features via a residual skip connection and processed through a Regressor Network to obtain frame-level importance scores. During training, a length regularization loss is applied, and during inference, the model selects key fragments to form a video summary within a specified duration limit.

The proposed video summarization methodology in [16] involves six primary phases: feature extraction, low-level refining, high-level refining, event clustering, static summary generation, and dynamic summary generation. It starts with the extraction of video and audio features frame by frame. Low-level refining uses The Structural Similarity Index (SSIM) and MFCC to remove redundant frames. High-level refining employs a CNN model to identify semantic differences between frames and refine the selection. Static summary generation can be done through event bucketing using SSIM or event clustering using K-means. Dynamic summary generation selects frames around candidate key frames for a smooth dynamic summary. MFCC features represent audio in audio extraction, and low-level refining utilizes SSIM and MFCC differences to discard redundant frames. High-level refining employs a CNN model (SqueezeNet) to distinguish frames within events. For the static summary, frames are either bucketed into events using SSIM or clustered with K-means based on VGG16 features. Dynamic summary selects frames around keyframes generated from static summary clusters. The approach is validated using the VSUMM dataset. This methodology offers a comprehensive approach to video summarization by considering both audio and video content, refining redundancies, and generating static and dynamic summaries. It leverages various techniques and features to improve summarization accuracy.

The authors in [17] discussed a novel approach called "Self-supervised Encoding Learning For Video Summarization (SELF-VS)" to address the challenge of limited annotated data for training models, especially in the video domain where collecting annotations is more costly than for images. The proposed method employs a Self-Supervised Learning (SSL) scheme based on a teacher-student framework. In this approach, a 3D CNN network trained for video classification serves as the teacher, providing supervision for a transformer-based encoder acting as the student. The intuition is that the pre-trained 3D CNN captures the semantics of entire videos, beneficial for video summarization, where the goal is to select important frames. The SSL scheme enables the student network to simultaneously learn frame features and importance scores. The video representation is obtained by aggregating intermediate frame features using scores, allowing the student network to prioritize important frames in video summarization. The SSL approach leverages knowledge distillation, aligning the teacher and student representations by minimizing cross-entropy. The fine-tuning phase follows standard supervised training, utilizing the pre-trained weights as a starting point and optimizing the network using mean squared error between ground truth and predicted scores. This two-phase approach, combining SSL pre-training and fine-tuning, addresses the challenge of limited annotated data for effective video summarization.

A novel approach called Multimodal Self-Supervised Progressive Video Summarization (SSPVS) introduced in [18]. The proposed approach utilizes a multimodal network trained in a self-supervised manner, capitalizing on semantic correlations between video and text. Video-text data are collected from the YouTube-based dataset named YTVT and are also utilized in the SumMe, TVSum, and OVP datasets. The multimodal network involves two unimodal encoders for text and visual information. The text encoder utilizes various types of text information from the Bidirectional Encoder Representations from Transformers (BERT), while the video encoder employs frame features from GoogLeNet. The learning objectives for self-supervised learning include modeling semantic correlations between video and text and capturing temporal dependencies within videos. The former involves coarse and fine-grained modeling, emphasizing both global and local information. These objectives involve fine-grained modeling of correspondence, as well as capturing temporal dependencies by recovering masked frames. The proposed progressive video summarization introduces a multi-stage approach for refining video sequences iteratively, with each stage enhancing input based on the previous one. The stages leverage pretrained video encoders to predict frame-level importance scores, employing a residual connection for stability. The final scores are computed by combining outputs from all stages. The framework is trained using ground truth scores, and summaries are generated by selecting shots to maximize the total score, constrained by a predefined summary length.

The proposed Video Summarization Generation Network Based on Graph Structure Reconstruction (VOGNet) in [19] aims to enhance video summarization by addressing redundancies in graph structures and learning structural features of videos. VOGNet comprises three key components: feature extraction, graph structure reconstruction, and feature fusion. In the feature extraction phase, video frames are processed to extract representative shot features, serving as nodes in a graph. The graph's edges are constructed based on the similarity between shots. The adjacency matrix is formed using cosine similarity, providing a representation of the video's structure. The graph structure is then reconstructed using the Variational Graph Autoencoder (VGAE), which employs a graph convolutional network for encoding. VGAE learns the low-dimensional vector representation of nodes and reconstructs the adjacency matrix. To optimize the adjacency matrix, Graph Attention Network (GAT) is introduced, considering weights of various neighbors and minimizing redundancy. The optimized attention scores are fused with shot depth features extracted by GoogLeNet. This fusion enhances the quality of shot selection. The combined features are input into a Multi-Layer Perceptron (MLP) for shot prediction scores. The loss function for VGAE involves reconstruction loss and KL divergence, evaluating the distance between generated and original graphs. VOGNet's overall loss function employs mean square error (MSE) to measure the gap between predicted and real shot scores.

The paper [20] presents a video summarization framework emphasizing amalgamating audio and visual data to enhance the summarization process. This framework comprises a Summarizer, which encompasses a frame selector, encoder, and decoder, and a Generative Adversarial Network (GAN) that includes a generator (decoder) and a discriminator. An underlying Variational AutoEncoder (VAE) is introduced to refine the training process further, thereby establishing an underlying video representation and introducing an additional frame scores vector. The proposed methodology introduces a keyframe selection mechanism that seeks to minimize the disparity between the characteristics of the original videos and the reconstructed videos derived from the predicted summaries. The summarizer and discriminator undergo unsupervised training until the discriminator can no longer distinguish between the reconstructed and original videos. The frame selector, encoder, and decoder employ LSTM models to perform their functions. The frame selector processes the frame features of the input video, yielding a normalized importance scores vector. Subsequently, the weighted frame features, resulting from the frame selector, are

Table 1
Summary of related works. This table presents an overview of notable video summarization publications, providing insights into the methods employed, features utilized, types of summaries generated, datasets employed for evaluation, and the corresponding results achieved.

Publication	Methods	Type of features	Summary type	Datasets	Performance	Year
[6]	Multi-CNN(AlexNet, GoogleNet, VGG16, IRv2), Sparse Auto Encoders (SAE), Random Forrest Classifier.	Visual	Static	VSUMM, OVP	F-score (VSUMM) = 83 F-score (OVP) = 82	2020
[14]	CNN(GoogleNet, ResNet50), Inflated 3D ConvNet(I3D), BiLSTM.	Visual	Dynamic	SumMe, TVSum	F-score (SumMe) = 54.4 F-score (TVSum) = 62.8	2021
[15]	CNN(GoogleNet), VGGish, BiLSTM, Self Attention Encoder(SAE).	Audio-Visual	Dynamic	SumMe, TVSum	F-score (SumMe) = 44.1 F-score (TVSum) = 59.7	2021
[9]	CNN(GoogleNet), RNN, Attention Mechanism.	Visual	Dynamic	SumMe, TVSum	F-score (SumMe) = 51.1 F-score (TVSum) = 61.4	2022
[16]	CNN(VGG16), MFCC, SqueezeNet, Clustering,SSIM.	Audio-Visual	Static, Dynamic	VSUMM	F-score (Static) = 58.03 F-score (Dynamic) = 27.1	2022
[17]	CNN(GoogleNet), 3D-CNN, KTS, Multi-Layer Perceptron (MLP), Self-Attention.	Visual	Dynamic	SumMe, TVSum	F-score (SumMe) = 39.8 F-score (TVSum) = 58.9	2023
[18]	CNN(GoogleNet), BERT, KTS, Residual Connection.	Visual	Dynamic	SumMe, TVSum YTVT, OVP	F-score (SumMe) = 48.7 F-score (TVSum) = 60.3	2023
[19]	CNN(GoogleNet), Graph Attention Network (GAT), Variational Graph Auto Encoders (VGAE), KTS, Multi-Layer Perceptron (MLP).	Visual	Dynamic	SumMe, TVSum	F-score (SumMe) = 49.8 F-score (TVSum) = 60.8	2023
[20]	CNN(GoogleNet), Bi-Directional LSTM, Self-Attention Modules, Variational AutoEncoder VAE.	Audio-Visual	Dynamic	SumMe, TVSum	F-score (SumMe) = 64.8 F-score (TVSum) = 63.1	2023
[21]	CNN(GoogleNet), VGGish, Bi-Modal Transformer, ShotConv Network.	Audio-Visual	Dynamic	SumMe, TVSum	F-score (SumMe) = 55.3 F-score (TVSum) = 69.3	2023

directed to the encoder, which generates a hidden state vector. The decoder, in turn, reconstructs a sequence of features that represent the input video, with the reconstructed sequence then subjected to assessment by an LSTM-based discriminator, which categorizes it as either ‘original’ or ‘summary’. The proposed SUM-GAN-AED model introduces an attention-based frame selector into the GAN architecture. This selector incorporates a self-attention layer, enabling it to capture long-term temporal dependencies more effectively than LSTMs. This transition leads to swifter computations and the production of more indicative frame selection scores. The paper also delves into variants of the model featuring self-attention layers, such as SUM-GAN-STD, SUM-GAN-ST, SUM-GAN-STSED, and SUM-GAN-SAT. In these variations, LSTMs are replaced with transformers at various stages, aiming to improve the modeling of temporal dependencies. These explorations provide insights into which components benefit most from enhanced temporal dependency modeling, thereby deepening our understanding of the architecture’s efficacy.

In [21], the authors introduce a Multimodal Hierarchical Shot-aware Convolutional Network (MHSCNet). The shot-aware network operates at the shot level, meaning its inner operations are defined based on shots. The proposed model leverages a CNN-pretrained model to extract visual features, VGGish for audio features, and a Bi-Modal transformer to generate video captions. MHSCNet effectively utilizes cross-modality information, incorporating audio and video captions in addition to the basic video frames, to create a robust frame-wise representation. This approach enhances the model’s understanding by considering various modalities. Furthermore, the hierarchical shot-aware convolutional network and the crossshot padding mechanism adapt shot-aware representations with both short-range and long-range temporal dependencies. This ensures that the model effectively captures temporal relationships at different scales within the video. Table 1 presents a summarize of the related works.

3. Proposed model

In this paper, we propose a deep audio-visual model for efficient dynamic video summarization, as illustrated in Fig. 1, with a multi-phase process detailed in Algorithm 1. The process begins with the Kernel-based Temporal Segmentation (KTS) [22] that divides the video into individual n shots and further splitting each shot into m frames and associated audios. The values of n and m vary from one video to another. Shot segmentation enhances the model focus on distinct visual events and facilitates comprehensive content representation. The proposed model consists of four main phases, namely feature extraction, audio-visual score prediction, shot selection and summarized video creation, through which the understanding of how the model operates will be apparent. These phases will be explained in detail through the following subsections.

3.1. Feature extraction phase

The feature extraction phase in the proposed deep audio-visual model plays a crucial role in capturing and representing the essential characteristics of both visual and audio content within the video. This phase serves as a foundation for subsequent phases in the summarization process, influencing the overall effectiveness of the model. Each frame f_j and its associated audio a_j within each shot are processed for feature extraction. Visual features are extracted using two pre-trained CNN models, “ResNet50” and “InceptionV3” with feature vectors denoted as $VF1$ and $VF2$. Early fusion combines these feature vectors to ensure they have the same shape, thus reducing overfitting and producing the final visual feature vector VF_{f_j}. Audio features are extracted using Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-Spectrogram representations, along with the VGGish model for audio representation, resulting in AF_{a_j}.

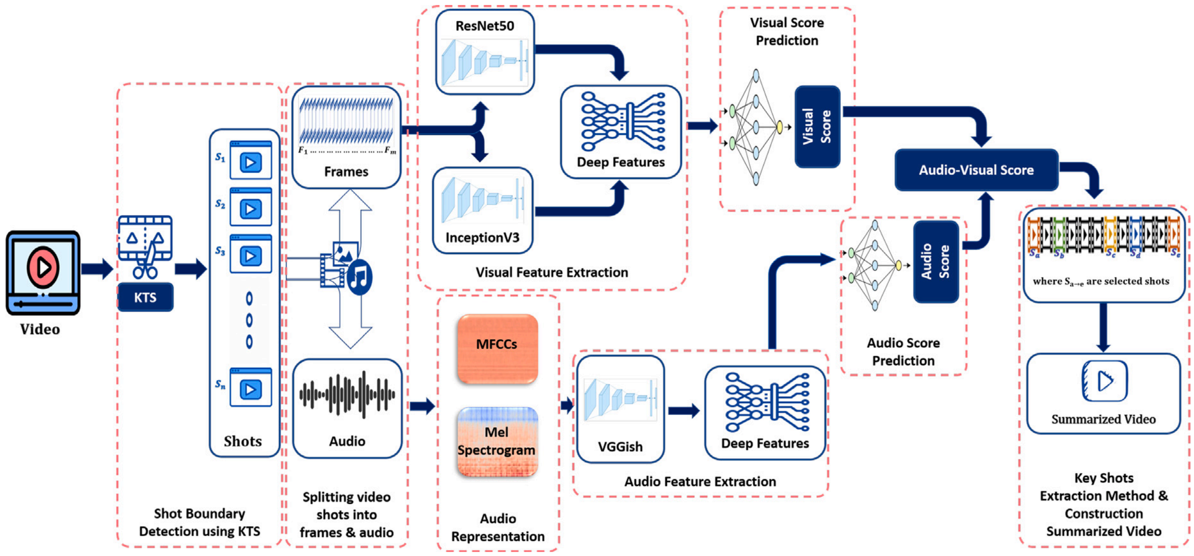

Fig. 1. The deep audio-visual model framework.

3.2. Audio-visual score prediction phase

The visual feature vector VF_{f_j} is passed through a maxpool layer and flattened, serving as input to an Artificial Neural Network (ANN) with a dense layer comprising 128 neurons. This ANN predicts the frame's visual importance score $V_{Score_{f_j}}$. The audio feature vector AF_{a_j} is also flattened and input into another ANN with dense layers of 2048, 1024, 2048, and 1024 neurons, followed by an output layer predicting the frame's audio importance score $A_{Score_{a_j}}$. The audio-visual score for each frame AV_{Score_j} is computed as the average of these scores, and for each shot $AV_{Score_{s_i}}$, it is the average of the frames' scores.

3.3. Shot selection phase

The model predicts the importance scores for each shot, and key shots are extracted based on a threshold value $Score_{th}$. The threshold is determined using Eq. (1), and if a shot's importance score meets the criteria defined in Eq. (2), it is selected and added to the $SelectedShots$ list.

$$Score_{th} = \frac{\sum_{j=1}^{n} AV_{Score_{s_j}}}{n} \tag{1}$$

$$\begin{cases} SelectedShots.append(s_j), & \text{if } AV_{Score_{s_j}} \geq Score_{th} \\ Discard(s_j), & \text{otherwise} \end{cases} \tag{2}$$

3.4. Summarized video creation phase

In the final phase, the dynamic summarized video DSV is constructed by assembling the selected shots from the $SelectedShots$ list. DSV's length is adjusted to be less than n while maintaining the same frame rate as the original video. It is natural for the DSV to be less than the n, because the n is the number of shots in the original video, and during the process of summarizing the video, some shots are neglected based on their lack of importance score.

4. Experimental results

In this paper, to advance dynamic video summarization, the outcomes of an efficient approach are unveiled, and the proposed model is to be tested against renowned benchmarks. This section introduces implementation details that would clarified model's configuration, pivotal datasets that have shaped our evaluation process and subsequently delves into the evaluation methods employed to gauge proposed model's efficacy.

Algorithm 1 The Deep Audio-Visual Model Algorithm.

```
Input: Video
Output: Dynamic Summarized Video ($DSV$)
Start
Cutting video into shots $\{s_1, s_2, ..., s_n\}$ using KTS
for $i = 1 : n$ do          ▷ where $n$ refer to total number of shots
    Splitting $s_i$ into $\{f_1, f_2, ..., f_m\}$ and audios $\{a_1, a_2, ..., a_m\}$
    for $j = 1 : m$ do          ▷ where $m$ refer to total no. frames in '$s_i$'
        $VF_{f_j}$ = ExtractDeepVisualFeatures($f_j$)
        $AF_{a_j}$ = ExtractDeepAudioFeatures($a_j$)
        $V_{Score_{f_j}}$ = PredictVisualScore($VF_{f_j}$)
        $A_{Score_{a_j}}$ = PredictAudioScore($AF_{a_j}$)
        $AV_{Score_j}$ = Average($V_{Score_{f_j}}$, $A_{Score_{a_j}}$)
    end for
    let $AV_{Score_{s_i}} = \frac{\sum_{j=1}^{m} AV_{Score_j}}{m}$
end for
let $SelectedShots$ be empty list []
for $i = 1 : n$ do
    if $AV_{Score_{s_i}} \geq Score_{th}$ then
        $SelectedShots.append(s_i)$
    end if
end for
Construct the dynamic summarized video '$DSV$' from $SelectedShots$
End
```

4.1. Implementation details

In order to reduce the number of redundant frames in our dataset, we downsampled all videos to two frames per second.

Shots are segments made up of sequences of frames, and within the same shot, the frames typically vary relatively steadily. Since our experiments require shot-level scores, we utilize the Kernel-based Temporal Segmentation (KTS) algorithm [22] to divide the video into video segments of varying lengths.

To ensure robust evaluation, the model is trained and evaluated on five different data splits for each dataset with an 80% train and 20% test split. The choice of k equal to 5 is aligned with the settings of state-of-the-art models for comparative purposes. Consistency in settings ensures a meaningful comparison with existing models in the field, we used the five splits provided in [23].

Fig. 2. Samples of SumMe videos.

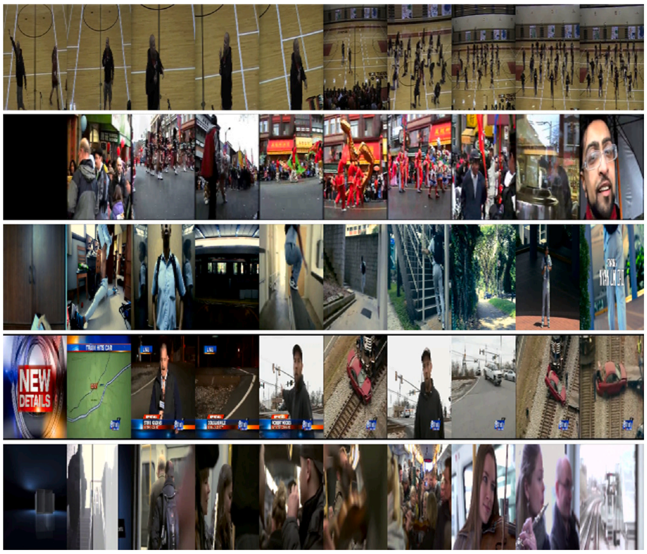

Fig. 3. Samples of TVSum videos.

The selection of neuron values in the Artificial Neural Network (ANN) architecture is a result of a systematic trial-and-error process. For predicting visual importance scores, the choice of 128 neurons is likely a result of empirical testing, aiming for a balance between model complexity and task performance. This size contributes to computational efficiency, crucial for large datasets or resource constraints. In the audio feature prediction section, the progression from larger (2048) to smaller (1024) layer sizes suggests a hierarchy of features, capturing intricate patterns in larger layers and refining information in smaller layers. The complexity of audio features may necessitate a larger initial representation, gradually reducing dimensionality for essential information. The threshold for shot selection, a crucial aspect, is dynamically determined during training or based on specific criteria, balancing precision and recall in shot selection.

4.2. Datasets

The proposed model used SumMe [24] and TVSum [25], renowned datasets in video summarization research, as benchmarks to evaluate and compare algorithms. Researchers utilize these datasets to evaluate their effectiveness and compare with state-of-the-art methods.

4.2.1. SumMe dataset

The SumMe dataset, a fundamental resource in video summarization studies, encompasses 25 videos from diverse domains like sports, news, and documentaries (refer to Fig. 2). Each video is associated with several human-generated summaries, varying in length and abstraction, offering crucial content. These summaries serve as vital benchmarks for assessing algorithm effectiveness in capturing key highlights and essential information. With its varied content genres, SumMe offers a comprehensive testing ground for video summarization models, addressing scenarios with distinct summarization requirements.

4.2.2. TVSum dataset

The TVSum dataset serves as a benchmark to validate video summarization, introducing distinct challenges compared to SumMe. Comprising 50 videos sourced from diverse TV shows (see Fig. 3), each video is paired with multiple human-annotated summaries. TVSum stands out due to its lengthier and more intricate summaries, presenting a heightened challenge for algorithms to efficiently distill essential content. it is a robust benchmark for assessing video summarization models across diverse content types.

4.3. Performance evaluation

The TVSum and SumMe datasets provides ground truth scores to each annotated frame. In the context of the proposed research, the ground truth refers to the annotated scores or importance levels assigned to individual frames in the videos within the datasets. Since the video split into n shots, the shot score will be the average score of these frames which construct it. So, let $GT_{Score_{s_i}}$ be the ground truth of shot i. The proposed model is evaluated by computing the F-Score based on the set of shots selected by the model. Let $GTSV$ ground truth summarized video that will be constructed from key shots vector $SelectedShots_{GT}$ by selecting shot s_i that ground truth importance score $GT_{Score_{s_i}}$ greater than threshold value $GT_{Score_{th}}$ as in Eq. (3) and Eq. (4). Now let overlapping shots between SV and $GTSV$ be $Overlap_{shots}$ as in Eq. (5)

$$GT_{Score_{th}} = \frac{\sum_{j=1}^{n} GT_{Score_{s_j}}}{n} \tag{3}$$

$$\begin{cases} SelectedShots_{GT}.append(s_j), & \text{if } GT_{Score_{s_j}} \geq GT_{Score_{th}} \\ Discard(s_j), & \text{otherwise} \end{cases} \tag{4}$$

$$Overlap_{shots} = \{shot : shot \in SV \bigcap GTSV\} \tag{5}$$

Eqs. (1) and (2) determine the threshold value for shot selection based on importance scores, calculating the mean and standard deviation of each shot. Eqs. (3) and (4) construct the ground truth summarized video (GTSV), setting a threshold based on mean and standard deviation, and selecting shots based on this threshold, ensuring the summary includes important shots based on annotated scores.

The metrics are calculated as given below. Let the summarized video SV consists of N_{SV} shots. Let $N_{Overlap}$ represents the shots in the generated summary that match the shots in the ground truth summary $GTSV$ and let N_{GTSV} be the shots in the ground truth summary, then

$$Precision = \frac{N_{Overlap}}{N_{SV}} \tag{6}$$

$$Recall = \frac{N_{Overlap}}{N_{GTSV}} \tag{7}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

The accuracy of the summary is estimated using the metrics Precision, Recall, and F-measure, which evaluate the temporal coherence

Table 2
The evaluation performance of proposed model for each fold on TVSum dataset. This table provides a detailed overview of the evaluation metrics for the proposed model on the TVSum dataset, measured across different folds, offering a comprehensive understanding of the model's summarization performance.

Fold	TVSum				
	Precision	Recall	F-score	Spearman's ρ	Kendall's τ
1	81.66%	85.09%	83.34%	82.22%	63.40%
2	78.12%	77.23%	77.67%	75.87%	57.71%
3	81.85%	81.10%	81.47%	81.15%	62.48%
4	72.04%	78.41%	75.09%	72.18%	53.97%
5	78.12%	79.80%	79.05%	80.80%	62.20%
Avg	78.40%	80.33%	79.33%	78.44%	59.95%

Table 3
The evaluation performance of proposed model for each fold on SumMe dataset. This table provides a detailed overview of the evaluation metrics for the proposed model on the SumMe dataset, measured across different folds, offering a comprehensive understanding of the model's summarization performance.

Fold	SumMe				
	Precision	Recall	F-score	Spearman's ρ	Kendall's τ
1	39.87%	62.71%	48.74%	25.90%	18.76%
2	65.08%	65.18%	65.13%	53.27%	40.83%
3	69.36%	71.75%	70.54%	64.21%	50.30%
4	74.18%	65.16%	69.38%	64.44%	51.21%
5	80.00%	80.23%	80.11%	76.48%	64.56%
Avg	65.70%	69.01%	66.78%	56.86%	44.53%

between the summarized video and the original, unrepeated video. However, it is crucial to acknowledge certain limitations. The metrics employed only partially consider the intricate nature of human preferences for video shots. While providing a thorough assessment of the summary quality, the evaluation may not fully capture the complexity of subjective preferences in video summarization. In order to alleviate this issue, proposed model uses Kendall's τ and Spearman's rank-based ρ metrics. These coefficients are used to evaluate the correlation between the predicted importance scores generated by the model and the ground truth importance scores. The proposed model is tested using k-fold cross-validation with k equal to five on the TVSum and SumMe datasets, as shown in Tables 2 and 3.

The difference in values between Tables 2 and 3 could be attributed to various factors, including the characteristics of the datasets, the complexity of video content, and the inherent challenges posed by SumMe and TVSum. The notable increase in performance from $k = 1$ to $k = 5$ in the SumMe Dataset could be attributed to several factors:

1. **Dataset Variability:** The SumMe dataset may have inherent variability, and different folds in cross-validation capture diverse subsets of the data. Averaging results over multiple folds might lead to a more representative evaluation.
2. **Randomness in Splitting:** The initial random split in k=1 might have resulted in a less favorable subset for evaluation. With k=5, the model is tested and validated on different subsets, providing a more comprehensive assessment.
3. **Model Stability:** Some machine learning models exhibit variability in performance depending on the data split. Multiple folds help assess the model's stability and generalization across different subsets of the data.

4.4. Comparison with state-of-arts

The 5-fold cross-validation was performed without shuffle splits, ensuring non-overlapping splits. This approach allowed all videos to be used in testing. Figs. 4 and 5 depict the F-score analysis for videos in TVSum and SumMe across the 5-fold validation. In SumMe, two videos lacked audio, which the proposed model leveraged to its advantage by using them during training for visual features. This strategy effectively turned this limitation into a benefit for the model. For more visualize results, Figs. 6, 7, 8 and 9 illustrate the actual score curve and prediction score curve for sample from these videos.

Table 4
The comparison of the proposed model's performance with the state-of-the-art techniques. This table presents a comprehensive comparison between the proposed model and state-of-the-art techniques in the field of video summarization. The evaluation is conducted based on F-score metrics over two benchmark datasets, TVSum and SumMe.

Methods	F-score	
	TVSUM	SUMME
DTR-GAN [26], 2019	61.3%	44.6%
SGSN [27], 2021	55.7%	41.5%
AC-SUM-GAN [28], 2021	60.6%	50.8%
ADSum [29], 2021	64.3%	46.1%
MSVA [14], 2021	62.8%	54.5%
AVRN [15], 2021	59.7%	44.1%
Bi-Convolutional-LSTM-GAN [30], 2022	71.6%	–
SELF-VS [17], 2023	58.9%	39.8%
SSPVS [18], 2023	60.3%	48.7%
VOGNet [19], 2023	60.8%	49.8%
MFST [31], 2023	73.7%	59.5%
SUM-GAN-AED [20], 2023	63.1%	64.8%
MHSCNET [21], 2023	69.3%	55.3%
Proposed model	**79.3%**	**66.7%**

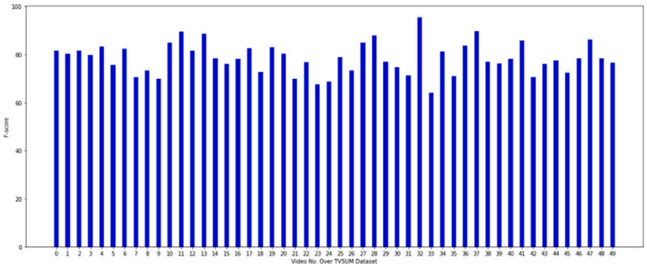

Fig. 4. The Visualization of F-scores for TVSum Videos. This figure visually represents the F-scores associated with TVSum videos, offering a clear and insightful depiction of the model's performance across different video samples.

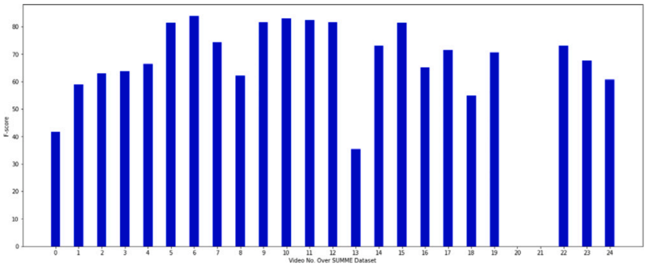

Fig. 5. The Visualization of F-scores for SumMe Videos. This figure illustrates the F-scores for SumMe videos, providing a comprehensive view of the proposed model's performance across various video samples. Note that videos #20 and #21, lacking audio, are excluded from visualization, emphasizing transparency about the limitations in capturing audio-related features for these specific videos.

The results in Table 4 demonstrate the performance of our proposed model compared to other state-of-the-art models using the same datasets with an 80% train and 20% test split. Our model outperformed all the compared models, achieving higher F-scores of 79.33% for TVSum and 66.78% for SumMe.

5. Conclusion

The research paper presents a dynamic video summarization approach that uses deep learning techniques to process audio and visual features. The model combines ResNet50 and InceptionV3 convolutional

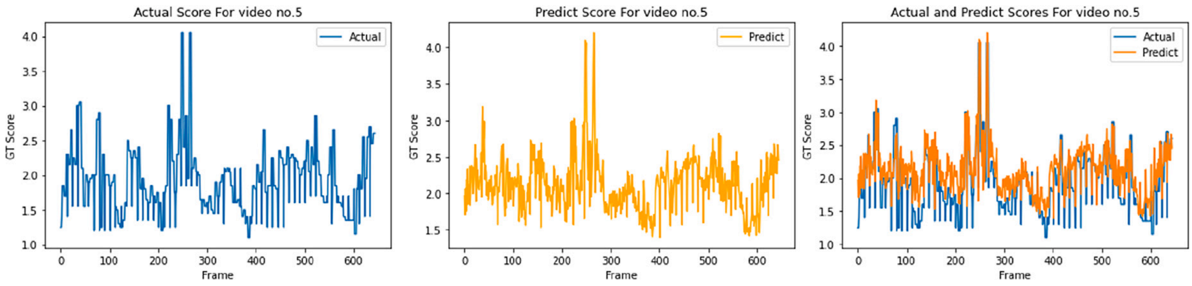

Fig. 6. The Visualization of Score Curves for 5th Video in TVSum. This figure illustrates the summarization process for TVSum's 5th video. The left shows the ground truth score curve, and the center displays the model-generated prediction score curve. The right illustrates their intersection, providing insights into the model's accuracy on a specific video.

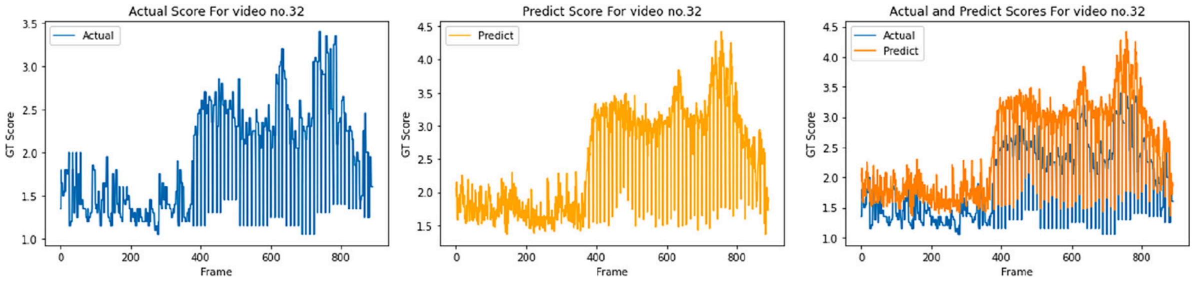

Fig. 7. The Visualization of Score Curves for 32th Video in TVSum. This figure presents the summarization process for the 32th video in TVSum. On the left is the ground truth score curve, while the center displays the model-generated prediction score curve. The right side illustrates the intersection between the two, offering a detailed assessment of the model's accuracy.

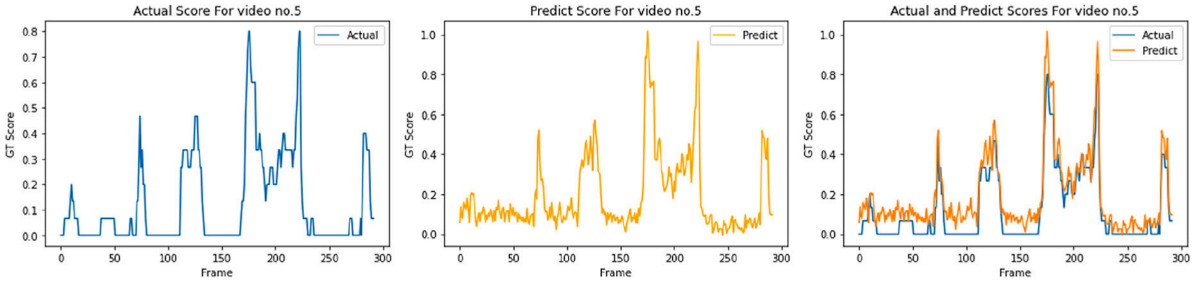

Fig. 8. The Visualization of Score Curves for 5th Video in SumMe. This figure illustrates the evaluation of the model for SumMe's 5th video: Ground truth curve (left), model's prediction curve (center), and their intersection (right), offering insight into model accuracy and capturing key moments.

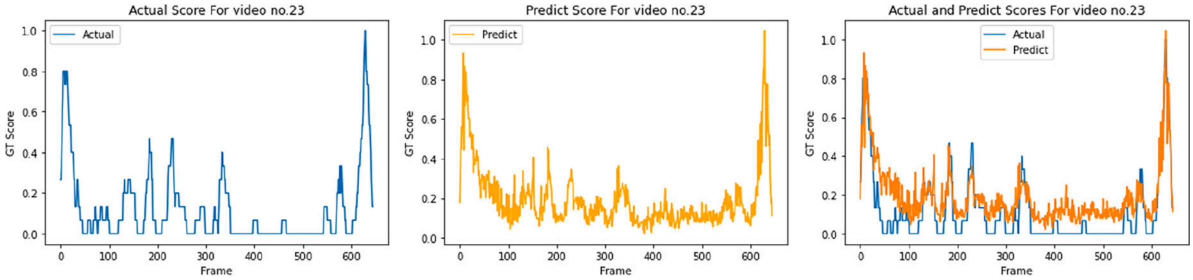

Fig. 9. The Visualization of Score Curves for 23th Video in SumMe. This figure presents score curves for SumMe's 23th video: Ground truth (left), model prediction (center), and their intersection (right), providing a detailed analysis of model accuracy and alignment with actual scores.

neural network (CNN) models for visual content features and employs MelFrequency Cepstral Coefficients (MFCCs), MelSpectrogram, and VGGish model for audio content. This comprehensive fusion allows for dynamic video summaries enriched with audio context, impacting video recommendation systems, content retrieval, and real-time video analysis. The model's performance is assessed using three metrics and compared to contemporary models on TVSum and SumMe datasets, demonstrating remarkable progress and superior efficiency. However, there are potential limitations in generalizing across various datasets, as the model's performance is influenced by specific datasets designed for video summarizing, which may be impacted by genres, styles, and settings of video material. Addressing these limitations is crucial to ensure the model's generalizability to handle various video footage. While the current research achieves significant milestones, several avenues for future exploration and enhancement emerge. These include investigating and implementing advanced techniques for richer, contextually nuanced video summaries, exploring the feasibility and optimization of the model for real-time video summarization applications, conducting user studies to assess the subjective quality of video summaries focusing on relevance and coherence, extending evaluation to diverse datasets for model robustness and generalization, and exploring opportunities for practical applications and collaborations with industry partners. These avenues for future work aim to push the boundaries of video summarization research, addressing both technical challenges and practical applications to contribute to the continued advancement of this dynamic field.

CRediT authorship contribution statement

Gamal El-Nagar: Software, Funding acquisition. **Ahmed El-Sawy:** Writing – original draft, Supervision. **Metwally Rashad:** Writing – review & editing, Writing – original draft, Supervision, Methodology.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors would like to thank the Deanship of Scientific Research, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia, for supporting this research via funding from Prince sattam bin Abdulaziz University, Saudi Arabia project number (PSAU/2024/R/1445).

References

[1] W.-l. Li, T. Zhang, X. Liu, A static video summarization approach via block-based self-motivated visual attention scoring mechanism, Int. J. Mach. Learn. Cybern. (2023) 1–12.

[2] H. Terbouche, M. Morel, M. Rodriguez, A. Othmani, Multi-annotation attention model for video summarization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 3143–3152.

[3] M. Sreeja, B.C. Kovoor, A multi-stage deep adversarial network for video summarization with knowledge distillation, J. Ambient Intell. Humaniz. Comput. 14 (8) (2023) 9823–9838.

[4] O. Issa, T. Shanableh, Static video summarization using video coding features with frame-level temporal subsampling and deep learning, Appl. Sci. 13 (10) (2023) 6065.

[5] A. Tonge, S.D. Thepade, S-VSUM: Static video content summarization using CNN, in: 2022 International Conference on Signal and Information Processing, (IConSIP), 2022, pp. 1–5, http://dx.doi.org/10.1109/ICoNSIP49665.2022.10007516.

[6] M.S. Nair, J. Mohan, Static video summarization using multi-CNN with sparse autoencoder and random forest classifier, Signal, Image Video Process. 15 (2020) 735–742.

[7] M. Abbasi, P. Saeedi, Adopting self-supervised learning into unsupervised video summarization through restorative score, in: Proceedings of IEEE International Conference on Image Processing, ICIP, IEEE, 2023.

[8] M.S. Nair, J. Mohan, VSMCNN-dynamic summarization of videos using salient features from multi-CNN model, J. Ambient Intell. Humaniz. Comput. 14 (10) (2023) 14071–14080.

[9] E. Apostolidis, G. Balaouras, V. Mezaris, I. Patras, Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames, in: Proceedings of the 2022 International Conference on Multimedia Retrieval, ICMR '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 407–415, http://dx.doi.org/10.1145/3512527.3531404.

[10] E. Apostolidis, E. Adamantidou, A.I. Metsai, V. Mezaris, I. Patras, AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization, IEEE Trans. Circuits Syst. Video Technol. 31 (8) (2020) 3278–3292.

[11] A. Singh, M. Kumar, Bayesian fuzzy clustering and deep CNN-based automatic video summarization, Multimedia Tools Appl. (2023) 1–38.

[12] B. Zhao, H. Li, X. Lu, X. Li, Reconstructive sequence-graph network for video summarization, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2021) 2793–2801.

[13] Y. Gao, N. Xu, X. Geng, Video summarization via label distributions dual-reward, in: IJCAI, 2021, pp. 2403–2409.

[14] J.A. Ghauri, S. Hakimov, R. Ewerth, Supervised video summarization via multiple feature sets with parallel attention, in: 2021 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2021, pp. 1–6s.

[15] B. Zhao, M. Gong, X. Li, Audiovisual video summarization, IEEE Trans. Neural Netw. Learn. Syst. (2021).

[16] M. Rhevanth, R. Ahmed, V. Shah, B.R. Mohan, Deep learning framework based on audio–Visual features for video summarization, in: Advanced Machine Intelligence and Signal Processing, Springer Nature Singapore, Singapore, 2022, pp. 229–243.

[17] H. Mokhtarabadi, K. Bahraman, M. HosseinZadeh, M. Eftekhari, SELF-VS: Self-supervised encoding learning for video summarization, 2023, arXiv preprint arXiv:2303.15993.

[18] H. Li, Q. Ke, M. Gong, T. Drummond, Progressive video summarization via multimodal self-supervised learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5584–5593.

[19] J. Zhang, G. Wu, S. Song, Video summarization generation based on graph structure reconstruction, Electronics 12 (23) (2023).

[20] M. Nektaria Minaidi, C. Papaioannou, A. Potamianos, Self-attention based generative adversarial networks for unsupervised video summarization, 2023, arXiv e-prints, arXiv–2307.

[21] W. Xu, R. Wang, X. Guo, S. Li, Q. Ma, Y. Zhao, S. Guo, Z. Zhu, J. Yan, MHSCNET: A multimodal hierarchical shot-aware convolutional network for video summarization, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.

[22] D. Potapov, M. Douze, Z. Harchaoui, C. Schmid, Category-Specific Video Summarization, vol. 8694, 2014, http://dx.doi.org/10.1007/978-3-319-10599-4_35.

[23] K. Zhang, W.-L. Chao, F. Sha, K. Grauman, Video summarization with long short-term memory, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 766–782.

[24] M. Gygli, H. Grabner, H. Riemenschneider, L. Van Gool, Creating summaries from user videos, in: ECCV, 2014.

[25] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, Tvsum: Summarizing web videos using titles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5179–5187.

[26] Y. Zhang, M. Kampffmeyer, X. Zhao, M. Tan, Dtr-gan: Dilated temporal relational adversarial network for video summarization, in: Proceedings of the ACM Turing Celebration Conference-China, 2019, pp. 1–6.

[27] Z. Li, L. Yang, Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3239–3247.

[28] E. Apostolidis, E. Adamantidou, A.I. Metsai, V. Mezaris, I. Patras, AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization, IEEE Trans. Circuits Syst. Video Technol. 31 (8) (2021) 3278–3292.

[29] Deep attentive video summarization with distribution consistency learning, IEEE Trans. Neural Netw. 32 (4) (2021) 1765–1775.

[30] M. Sreeja, B.C. Kovoor, A multi-stage deep adversarial network for video summarization with knowledge distillation, J. Ambient Intell. Humaniz. Comput. 14 (8) (2022) 9823–9838.

[31] J. Park, K. Kwoun, C. Lee, H. Lim, Multimodal frame-scoring transformer for video summarization, 2023, arXiv preprint arXiv:2207.01814.