A Comprehensive Survey on Video Summarization: Challenges and Advances

Hongxi Li, Yubo Zhu, Zirui Shang, Ziyi Wang, Xinxiao Wu, *Member, IEEE*

***Abstract*—Video data is growing exponentially daily due to the popularity of video-sharing platforms and the proliferation of video capture devices. The video summarization task has been proposed to remove redundancy while maintaining as many critical parts of the video as possible so that users can browse and process videos more effectively, which has received increasing attention from researchers. The existing research addresses the challenges faced by video summarization methods from various perspectives, such as temporal dependency, data scarcity, user preference, and high precision. This paper reviews representative and state-of-the-art methods, analyzes recent research advances, datasets, and performance evaluations, and discusses future directions. We hope this survey can help future research explore the potential directions of video summarization methods.**

***Index Terms*—Video summarization; Deep learning; Weakly-supervised video summarization; Unsupervised video summarization; Query-focused video summarization**

I. Introduction

In recent years, the proliferation of video recording and sharing platforms has engendered an exponential growth of video data across diverse domains [1], [2], including entertainment, education, sports, and news. Face with massive amounts of videos, the tasks of storage, retrieval, and browsing require a lot of resources. At the same time, meaningless or redundant video content will make the video processing cumbersome and time-consuming [3]. To address these challenges, video summarization (VS) has emerged as a key solution. This method aims to extract the essence of a video by retaining the most salient content and removing irrelevant information. The process involves inputting an original video and generating a summary video containing key information, thereby improving efficiency and user engagement.

The essence of summarization is to transform a large amount of redundant content into brief content that can cover the original information. This transformation can occur in the same modality or across different modalities. Summarization methods can be divided into different fields according to the modality of information, including text summarization [4], multimodal sentence summarization [5]–[7], multimodal abstractive summarization [8], and multimodal product summarization [9], [10]. Different from video summarization, the above-mentioned summarization techniques generate abstractive summaries where the output content does not exist in the original input, and they all produce textual output from multimodal inputs. In contrast, the video summary content is represented by key parts of the original video, referred to as extractive summarization. The differences between the aforementioned summarizations are illustrated in Figure 1.

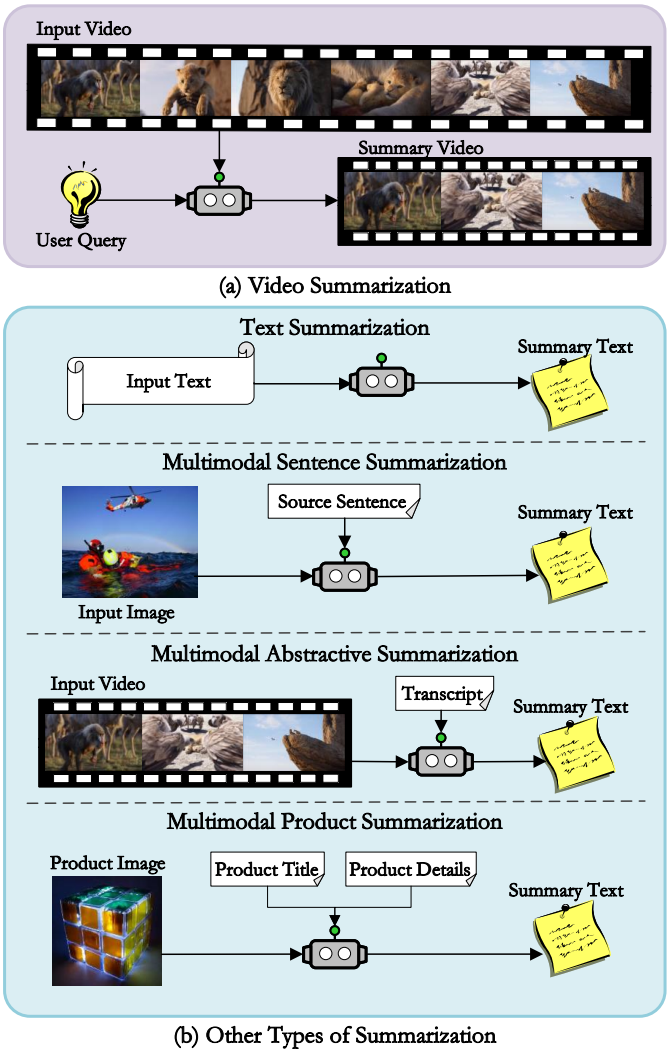


Fig. 1. Illustration of video summarization and other types of summarization.

Video summarization has attracted a lot of attention in the academic community and has made remarkable progress. These advances include the transitions from storyboard [11] to video skimming [12], from single-view to multi-view [13],

Hongxi Li, Yubo Zhu, Zirui Shang, and Ziyi Wang are with the Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: lihongxi@bit.edu.cn; 3120211052@bit.edu.cn; shangzirui@bit.edu.cn; ziyiwang@bit.edu.cn).

Xinxiao Wu is with the Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China, and also with the Guangdong Provincial Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China (e-mail: wuxinxiao@bit.edu.cn).

TABLE I
COMPARATIVE ANALYSIS OF SURVEY OR REVIEW RAPERS ON VIDEO SUMMARIZATION. FIR.: FIRST-PERSON VIEW VIDEOS, MUL.: MULTI-VIEW VIDEOS, PER.: PERSONALIZED SUMMARIES, COM.: COMPRESSED DOMAIN VIDEOS, SUR.: SURVEILLANCE VIDEOS, MED.: MEDICAL VIDEOS, SPO.: SPORTS VIDEOS.

Paper	Year	Criteria of Taxonomy	Application Domain						
			Fir.	Mul.	Per.	Com.	Sur.	Med.	Spo.
Molino *et al.* [25]	2016	Type of summary.	✓						
Kini *et al.* [26]	2019	Type of summary.		✓	✓	✓		✓	✓
Sen *et al.* [27]	2019	Numbers of views; Mechine learning method.							✓
Basavarajaiah *et al.* [28]	2019	Type of summary; Domain; Machine learning method; Feature; Information source; Data-dimension.				✓			
Haq *et al.* [29]	2020	Feature.					✓		✓
Raut *et al.* [30]	2020	Type of summary; Feature.						✓	
Hussain *et al.* [31]	2021	Feature.		✓					
Apostolidis *et al.* [32]	2021	Machine learning method.							✓
Vasudevan *et al.* [33]	2021	Feature.							✓
Tiwari *et al.* [34]	2021	Type of summary; Information source.	✓	✓	✓		✓		✓
Parihar *et al.* [35]	2021	Machine learning method.							✓
Narwal *et al.* [36]	2022	Type of summary; Numbers of views; Information source.	✓				✓		✓
Raval *et al.* [37]	2022	Feature.							✓
Moussaoui *et al.* [38]	2023	Feature; Machine learning method.	✓						✓
Correia *et al.* [39]	2023	Type of summary.						✓	
Meena *et al.* [40]	2023	Data-dimension; Numbers of views.			✓		✓		✓
Sabha *et al.* [41]	2023	Type of summary; Numbers of views; Feature; Time of summary.	✓				✓	✓	✓
Shambharkar *et al.* [42]	2023	Type of summary; Domain; Information source; Machine learning method.			✓	✓	✓		✓
Peronikolis *et al.* [43]	2024	Type of summary; Feature; Domain; Information source; Machine learning method; Time of summary.	✓		✓				✓
This paper	2024	Challenging issues of video summarization.	✓	✓	✓	✓	✓	✓	✓

from general summary to query-focused summary [14], and from conventional video to compressed videos [15]. Research efforts have focused not only on videos in daily life but also on various specialized fields, such as sports analysis (including soccer [16], [17] and cricket [18], [19]), surveillance [20]–[22], medical diagnosis (including bronchoscopy procedure [23] and wireless capsule endoscopy [24]). Sports video summarization generates a highlight reel of a game for viewers. Surveillance video summarization aids in identifying security incidents within extensive surveillance videos. Medical video summarization helps to obtain video captured by medical equipment that is used for diagnostic purposes.

Existing surveys on video summarization have predominantly employed taxonomies based on technical characteristics such as summary type (storyboard vs. video skimming), feature modalities (event-based vs. motion-based), machine learning approaches (supervised vs. unsupervised), or application domains (pixel vs. compressed domain), as systematically compared in Table I. While these categorizations—spanning information sources (internal/external/hybrid), view multiplicity (single/multi-view), data dimensions (2D/3D), and temporal aspects (real-time/static)—provide methodological overviews, they largely fail to address fundamental research challenges. Although recent works [28], [42], [43] have incorporated domain-specific analyses (e.g., medical, sports) and emerging paradigms like multi-view summarization, they lack a unified framework connecting these dimensions to core obstacles such as data scarcity, temporal dependency, domain adaptation, and user preference integration. Critical gaps persist in addressing real-time constraints and evaluation without ground truth. In contrast, our challenge-driven taxonomy synthesizes existing classifications while explicitly highlighting unresolved problems, shifting the field from descriptive technical categorizations toward problem-oriented solutions that bridge methodological and practical research needs.

We review and categorize the existing VS methods, from a novel perspective of the challenges faced by VS, including addressing temporal dependencies, handling data scarcity, obtaining high precision, and meeting user preferences. Additionally, this paper covers VS methods in several active domains: first-view, multi-view, personalized, compressed-domain, surveillance, medical, and sports video summarization.

The main contributions of this paper are as follows:

- We present a novel taxonomy based on the challenging problems faced by video summarization.
- We present a comprehensive survey on video summarization methods in a wide range of active domains.
- We conduct an in-depth analysis of the state-of-the-art video summarization methods through qualitative and quantitative results.
- We explore several opening research challenges for the future development of video summarization.

The remainder of this paper is organized as follows. Section II presents a general framework of video summarization. Section III systematically reviews related work based on the introduced taxonomy. Section IV-A describes existing datasets of video summarization. Section IV-B presents evaluation metrics, and Section IV-C analyzes comparison results between existing methods. Section V discusses future directions. The conclusion is provided in Section VI.

II. Video Summarization

The goal of video summarization is to generate a brief and representative target video from the input video. Depending on the form of the output video, video summaries can be roughly categorized into three types, including static video summary, dynamic video summary, and query-based video summary, as shown in Figure 3. Static video summary is represented by a subset of the frame sequence from the input video, with frames being discontinuous from each other. Dynamic video summary is represented by a subset of shot sequences obtained after shot segmentation of the input video, where frames within the same shot are continuous. Query-based video summary, a variant of dynamic video summary, generates different subsets of shot sequences in response to various user queries.

Video summarization can be formulated as a subset selection problem, which selects key-frames from the frame set in the input video as the summary video. The pipeline of video summarization methods generally includes feature extraction, evaluation, and summary generation. The general framework is shown in Figure 2.

1) **Feature Extraction:** Feature extraction aims to transform the input video and user query into numerical features that can be used for generating summaries. Deep learning-based video summarization methods typically employ Convolutional Neural Networks (CNNs) to extract visual features from videos. Some studies use 2D convolutions (e.g., VGG [44], GoogleNet [45], ResNet [46]) to extract features frame by frame, while others employ 3D convolutions [47] to capture temporal information. For query-based video summarization, the encoders corresponding to each modality are required. To maintain the coherence of the summary video, video skimming usually involves extracting shot-level features, which requires segmenting the shots through typical algorithms such as uniform segmentation and kernel temporal segmentation (KTS) [48].
2) **Evaluation:** The extracted features from video frames are then evaluated to predict the importance scores of these frames. For query-based video summarization, the importance scores reflect both the representativeness of the video frames to the original content and their relevance to the user query. Typically, the evaluation is implemented using networks based on CNNs, Recurrent Neural Networks (RNNs) [49], or attention mechanisms [50].
3) **Summary Generation:** Given the importance scores of video frames, the summary generation process requires a subset selection strategy to identify the most critical parts of the original video. Considering the need to select the most important segments while adhering to video length constraints, most studies use the 0/1 knapsack algorithm [51] for key frame selection.

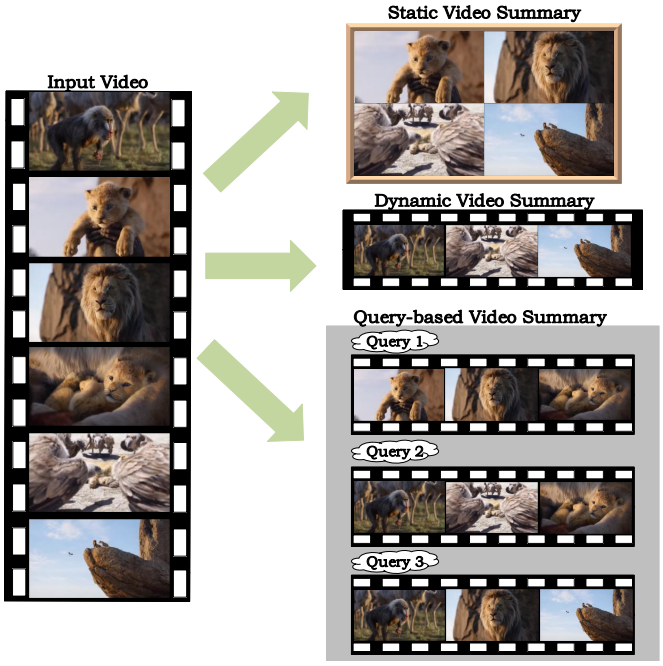


Fig. 3. Different types of video summary.

To be more specific, the general video summary generation process is detailed as follows. Let Q represent the user query and $V = [v_1, v_2, ..., v_N]$ denote the input video with N frames (assuming no frame sampling is performed). The feature extractors process both the user query and the input video, yielding the query feature q and the video frame feature $F = [f_1, f_2, ..., f_N]$. These extracted features are then used by the evaluation network to compute importance scores and by the summarization generation algorithm to produce a summary video, as described by the following equations:

$$I = \pi(f_1, f_2, ..., f_N|q) \tag{1}$$

$$S = Knap(I) \tag{2}$$

where π denotes the evaluation network, I denotes the importance score with $I = [i_1, i_2, ..., i_N]$, and S denotes the summary video with $S = [s_1, s_2, ..., s_k], s_i \in V, k < N$.

III. Taxonomy

Based on the challenges encountered in video summarization, which are also being addressed by existing methods, we

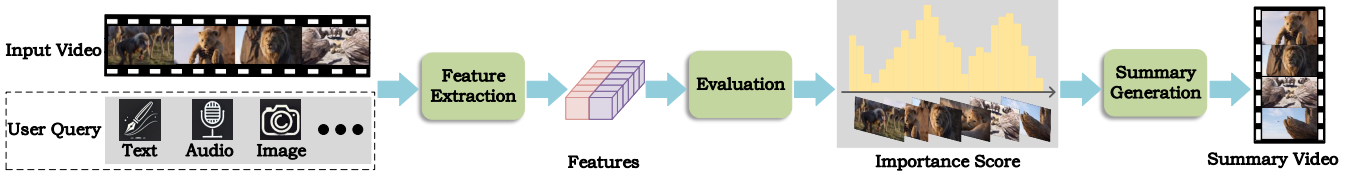


Fig. 2. A general framework of video summarization methods. User query serves as a a component of the input for query-based video summarization.

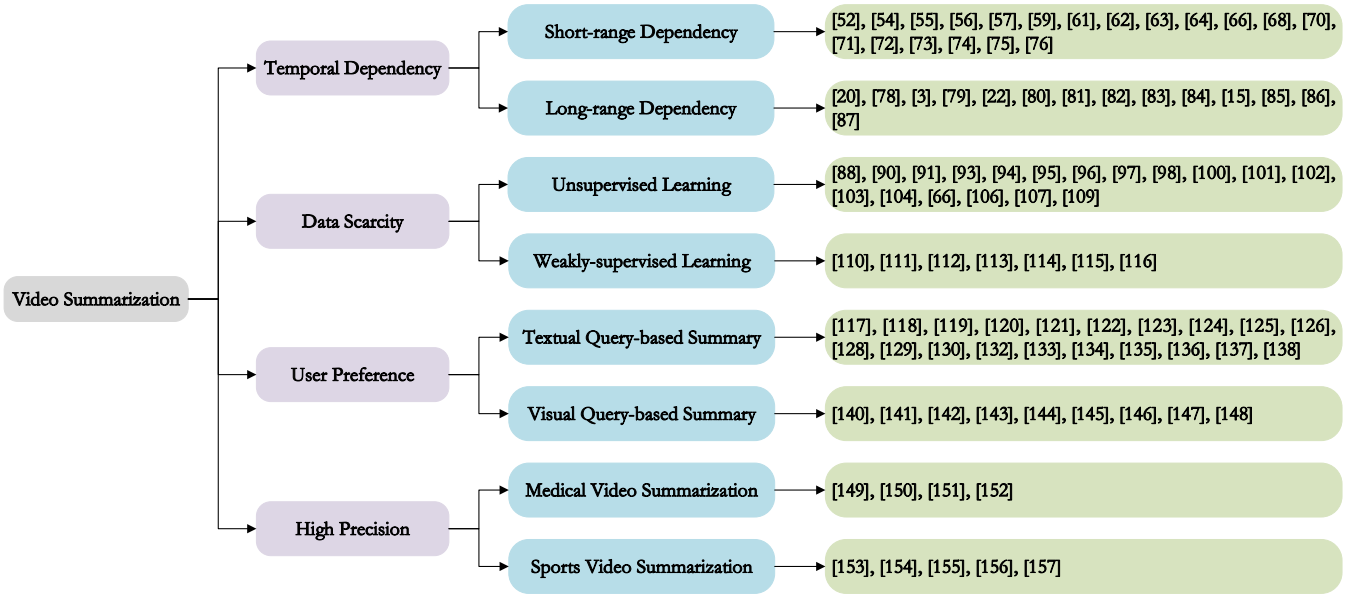


Fig. 4. A taxonomy of video summarization methods based on the challenges they face.

categorize the existing video summarization methods into addressing temporal dependency, data scarcity, user preferences, and high precision, as shown in Figure 4.

A. Temporal dependency

Modeling the temporal dependence of short-range and long-range relationships between video frames is a challenging problem for Video Summarization (VS) models, because summarization algorithms that rely solely on visual cues without considering temporal dependencies will incorrectly eliminate important frames.

1) Short-range temporal dependency: Modeling the temporal dependence of short-range relationships helps to capture the dynamic changes and connections between adjacent frames and effectively represent immediate transitions and actions in the video, ensuring that the summary accurately reflects the flow and key events of the video.

LSTM is particularly advantageous in modeling short-range temporal dependency. Zhang *et al.* (2016) [52] first apply LSTM and its variants to VS. They use forward and backward chains in bidirectional LSTM [53] to model past and future temporal dependencies. However, Yue *et al.* (2015) [54] find that LSTM has limited ability in modeling dependencies, and its applicable video length is not more than 80 frames, while the original video in VS usually lasts longer. To address this issue, Zhao *et al.* (2017) [55] design a hierarchical RNN with two layers. The first layer uses LSTM to encode video shots captured from the original video, and inputs the final hidden state of each shot to the next layer to establish the short-range dependency. The second layer uses Bi-LSTM to establish the long-range dependency and calculates the confidence of each shot as a key shot. To model 3D contexts, Huang *et al.* (2018) [56] employ a divide-and-conquer strategy, which first extracts features from video frames using a two-stream VGG model, then captures the short-term temporal context using 1D convolution, and finally captures the global temporal context using LSTM.

Furthermore, Zhao *et al.* (2021) [57] try to use models other than LSTM to deal with the dependency problem. They first use the KTS algorithm to split the video into shots. For the frame-level sequence in each shot, Bi-LSTM is used to capture the temporal dependency between frames and obtain shot-level features. For the shot-level dependency, a graph convolutional network (GCN) [58] is constructed with shots as nodes and the differences between shots as edge weights to capture the temporal dependency between different shots. Zhu *et al.* (2022) [59] use dynamic graph modeling to learn the spatio-temporal representation of videos. More specifically, they first adopt R-CNN [60] to extract object-level spatial information in video frames, then use the spatial information as nodes to construct a spatial graph, and finally use the spatial graphs corresponding to different frames as nodes to construct a time graph. Ji *et al.* (2019) [61] use LSTM as an encoder to obtain representations and apply the attention mechanism to the decoder to calculate the weights of different frames. In this way, the decoder can capture the context information over a longer time range based on attention weights when predicting importance scores. Khan *et al.* (2024) [62] proposes a deep pyramidal refinement network to extract and refine multi-scale progressive features. Jin *et al.* (2024) [63] develops a multiscale representation scheme to capture complex temporal relationships by incorporating the U-Net architecture and local attention into the encoder-decoder structure.

The above-mentioned methods usually use RNN represented by LSTM to generate video summaries, but these LSTM-based models have inherent limitations. The calculations in LSTM are usually from left to right, which means that one frame must be processed at a time, and each frame must wait until the previous frame to be processed. Due to this sequential nature, the calculations in LSTM cannot be easily parallelized to take full advantage of GPU hardware.

To address this problem, Rochan *et al.* (2018) [64] replace the spatial convolution in the Fully Convolutional Networks (FCN) [65] commonly used in semantic segmentation with temporal convolution, thus improving the parallelism by using temporal convolution to process all frame-level features in the video simultaneously.In a more recent method, Liu *et al.* (2022) [66] further explore FCN for VS. They use a 3D feature extraction network that computes n-frame features at a time to extract spatio-temporal features [67]. In addition, their 3D spatio-temporal U-Net (3DST-UNet) takes spatio-temporal video features from the previous spatio-temporal 3D CNN as input, which has the advantage of being able to process the entire video sequence simultaneously. Alharbi *et al.* (2024) [68] incorporates a fine-tuning InceptionV3 [69] backbone for feature extraction and introduces an encoder-decoder framework containing convolution and channel attention for capturing complex relationships. Singh *et al.* (2024) [70] selects representative frames using Bayesian fuzzy clustering (BFC) and refines those frames using deep CNN.

Moreover, the attention mechanism can also be used to solve the problem of low efficiency caused by the need for LSTM to process frame-level features cyclically or sequentially. Fajt *et al.* (2018) [71] first propose a pure attention network VASNet for VS. Since the vanilla attention model has equal access to all past and future inputs, VASNet takes the features of all frames as a matrix and inputs it into the attention network at one time, which has higher computational efficiency than the linear processing characteristics of LSTM. Later, Zhu *et al.* (2020) [72] compare different backbone models (LSTM, Bi-LSTM, GCN, Attention), which shows that the pure attention mechanism can generate high-quality video summaries while improving parallelism, and even has advantages over traditional LSTM and GCN. Zhang *et al.* (2024) [73] models global and local contextual information in a parallel manner and adaptively fuses features with semantic scale differences. Hsu *et al.* (2023) [74] introduces a transformer-based network for the first time in video summarization, capturing both temporal and spatial relationships simultaneously. Terbouche *et al.* (2023) [75] makes use of labels annotated by multiple human annotators to optimize a temporal attention-based network. Zhong *et al.* (2023) [76] utilizes attention alignment to extend representations from a single semantic feature space to a dual feature space of semantics and vision while promoting learning of higher-level dual features through Graph Information Bottleneck [77].

2) Long-range temporal dependency: Establishing long-range temporal dependencies in long videos is crucial for capturing the overall context and key events, but it poses challenges such as high computational complexity, information redundancy, and large storage and memory requirements.

In surveillance video summarization, the videos are typically continuous, long-duration, and uninterrupted, capturing a wide range of activities from routine actions to unexpected events. To create meaningful summaries, models must identify and correlate key events and behaviors that occur over a long period of time, which requires effectively capturing and processing long-range temporal dependencies. Zhang *et al.* (2016) [20] adopts sparse coding with generalized sparse group lasso to update online a dictionary of video features and a dictionary of spatio-temporal feature correlation graphs. Song *et al.* (2016) [78] propose to combine the Random Forest and the trajectory features to detect abnormal events, and design a disjoint max-coverage algorithm to generate a summarized sequence with maximum coverage of interested events and minimum number of frames. Panda *et al.* (2017) [3] propose to capture multiview correlations via through embedding to extract a diverse set of representatives, and model the sparsity when selecting representative shots for summarization. Muhammad *et al.* (2019) [79] use image memorability predicted from a fine-tuned CNN model, along with aesthetic and entropy features to keep the summary interesting and diverse, and propose a hierarchical weighted fusion mechanism to generate aggregated scores for segmented shots. Later, they [22] propose to use deep features extracted from CNNs for shot segmentation and use image memorability and entropy to select keyframes to improve the efficiency and intelligence.

More recently, Mahum *et al.* (2023) [80] obtain silhouette images of surveillance videos and apply Zernike Moments and R-Transform to construct combined features for key-frame extraction. Xu *et al.* (2021) [81] designs a reward function for calculating similarity and difference in crowd location and density to measure the set of different crowd behaviors. Sharma *et al.* (2023) [82] uses a new Deep LSTM model to extract the features from all segments of frames and cluster them using k-means clustering. Then, such features are trained using a pairwise deep learning model to get highlight scores as well as leverage the comparative similarity among pairs of the highlight as well as non-highlight segments. Ren *et al.* (2024) [83] uses sliding windows to create an indefinite-length video token sequence dynamically and explicitly binds visual content with each frame's timestamp. Papalampidi *et al.* (2024) [84] use pre-trained models, Video ViT and BERT, as the visual and textual encoders, respectively. They pre-train the model using Noise Contrastive Estimation (NCE) loss, which helps the model capture long-range visual dependencies in videos, such as the continuity and causal relationships of actions.

In compressed-domain video summarization, compressed videos only retain sparse information such as key-frames and motion vectors, leading to loss of temporal and spatial details. Critical events may span multiple key-frames, and the lack of intermediate details makes it difficult for models to accurately detect and correlate these events over a long time span, and then effectively capture and handle long-range dependencies. Almeida *et al.* (2013) [15] extract color histograms from the HSV color space as the compressed video features and introduce a zero-mean normalized cross-correlation metric to detect groups of video frames with similar content and to select a representative frame for each group. Hernandez *et al.* (2016) [85] propose a color descriptor combined with scene detection strategies to effectively detect gradual and abrupt transitions. Basavarajaiah *et al.* (2018) [86] extracts frame features from just the I-frames of the video, and classifies them into a predefined number of classes using K-means clustering. Then, the frame which is located at the border of a class in

the sequential order is selected to be included in the summary. Fei *et al.* (2018) [87] firstly segment a long video into shots according to DC-based mutual information, then select optimal object outliers from each shot using importance ranking and optimization, and finally employ an improved KNN matting method to seamlessly splice these outliers into the final key-frames.

B. Data scarcity

Supervised learning requires sufficient data, but the annotated data of VS task is relatively difficult to obtain. There are three main reasons: firstly, the annotator needs to watch the original video from beginning to end, which is an expensive and time-consuming process, especially when the video is long; secondly, different annotators are subjective about the representative frames in the same video; thirdly, it is difficult to obtain annotations for the videos in some specific domains such hospital and military domains. To address this challenge, some methods introduce unsupervised learning and weakly-supervised learning.

1) Unsupervised learning: Unsupervised methods aim to learn the intrinsic properties of VS. According to their properties, unsupervised methods can be roughly divided into two categories: learning representativeness by generative adversarial networks and learning other specific properties through different hand-crafted reward functions.

Unsupervised learning lacks ground-truth annotations to guide the generation of video summaries. Under this premise, an intuitive way to obtain a representative summary is to reduce the distance between the summary and the original video. The main idea of adversarial learning is that the generator first obtains the summary video based on the importance score of the original video sequence, then reconstructs the summary video into a video of the same length as the original video, and finally hands it over to the discriminator to distinguish the original video from the reconstructed video. Mahasseni *et al.* (2017) [88] first propose an adversarial-based unsupervised VS. In their SUM-GAN framework, the generator consisting of three LSTMs selects a subset of the original video, and reconstruct the subset into a feature sequence of the same length as the original video sequence using the reconstruction loss in Variational Autoencoder (VAE) [89]. Then, a discriminator (i.e., LSTM) is used to distinguish the reconstructed sequence from the original sequence.

Considering the problems that the mapping from the original video to the summary in SUM-GAN suffers from information loss and fails to capture the long-range temporal dependency of the entire video, He *et al.* (2019) [90] map the original frame features to the weighted frame features in the generator to avoid sparse mapping, and introduce a self-attention module to capture the long-range temporal dependency between frames. Jung *et al.* (2019) [91] point out another problem in SUM-GAN. When the importance scores are uniformly distributed, the original video can be reconstructed by the original features, making it difficult to learn discriminative features to find critical shots. To this end, they propose a variance loss to prevent the importance scores from being flattened. In addition, in order to learn discriminative features in long videos, they design a two-stream network to process local and global features.

With the further development of GAN and the widespread concern of attention mechanism, Cycle-GAN [92] and attention mechanisms have also been used for unsupervised VS based on adversarial learning. Yuan *et al.* (2019) [93] use two pairs of generators and discriminators to maximize the mutual information between the summary and the original video. The forward discriminator is used to distinguish between the summary features and the original features reconstructed from the summary features. The backward discriminator is used to distinguish between the original features and the summary features reconstructed from the original features. This cyclic processing helps to enhance the consistency between the original and the summary video. Apostolidis *et al.* (2020) [94] introduce the attention mechanism to VAE in two ways. One is to use the output of the encoder and the attention value to generate a decoded representation of the input sequence. The other is to use the output of the encoder and the previous hidden state of the decoder to calculate the attention of the next frame. The performance of the former is worse than that of the latter due to the difficulty in efficiently defining two latent spaces in parallel to the continuous updating of the model's components during training. Minaidi *et al.* (2023) [95] uses a self-attention mechanism for frame selection, combined with LSTMs for encoding and decoding based on GAN. Yu *et al.* (2024) [96] introduces the diffusion model to avoid the training instability problem caused by GAN's alternate training generator and discriminator. Abbasi *et al.* (2023) [97] proposes to reconstruct an entire video from a given summary and converts the reconstruction loss values into frame importance scores.

The above adversarial learning-based methods mainly focus on the reconstruction ability of video summaries. A high-quality summary not only has good reconstruction ability, but also has diversity that makes the abstract more attractive. Based on the above observations, reinforcement learning has been proposed to learn specific properties of VS.

The reward function is the core of reinforcement learning. Zhou *et al.* (2018) [98] propose a diversity-representativeness reward function that jointly accounts for the diversity and representativeness of the generated summaries. The proposed reward consists of a diversity reward and a representative reward. The former measures the dissimilarity of the selected frames at a certain time distance, and the latter formulates the degree of representativeness of a video summary as the k-medoids problem [99] to minimize the mean square error between the selected frame and its neighboring frames. Compared with diversity and representativeness, the discriminability of video actions can also be used as a reward function for reinforcement learning. Lei *et al.* (2018) [100] propose an action parsing-driven VS model based on reinforcement learning, which holds that analyzing actions in videos is conducive to mining the semantic information contained in videos. Wang *et al.* (2024) [101] optimize the model alternatively using a horizontal policy for measuring the amount of information conveyed by summaries or a vertical policy

for measuring the representativeness and diversity of each frame. Pang *et al.* (2023) [102] propose three metrics featuring a desirable keyframe: local dissimilarity, global consistency, and uniqueness by leveraging the contrastive losses. Zang *et al.* (2023) [103] focuses on the dependencies between any arbitrary frames while ignoring the redundant distance noises between two frames.

In addition to the reward function, the modification of the summary model itself is also a research direction. In previous unsupervised methods, LSTM is mainly used as a decoder to generate summaries, but the use of sigmoid and hyperbolic activation functions in LSTM results in gradient attenuation with layers. In order to alleviate the problem of attenuation gradient, Yaliniz *et al.* (2021) [104] use an independent recurrent neural network(IndRNN) [105] as a decoder. Compared with the RNN gradient involving the Jacobian matrix of the activation function, the IndRNN gradient only involves the exponential terms of the recurrent weight, which is easier to regulate and limited to a certain range, making it possible to train a deeper model. Liu *et al.* (2022) [66] change the convolutional network commonly used in the encoder to a 3D convolutional network, and use 3D U-Net as a decoder to better model the spatial and temporal relationships in video sequence.

Song *et al.* (2015) [106] attempts to use video captions for unsupervised learning. They proposed a title-based VS using video captions to find visually important shots, approximating video frames and images searched using the title through geometrical constraints. Considering the long-range dependency on the time axis in long videos, Jung *et al.* (2020) [107] add relative position representation to the scaled dot-product attention (RPE) [108]. Combining adversarial learning and reinforcement learning is also an effective method for unsupervised VS. Apostolidis *et al.* (2020) [109] embed an Actor-Critic model into the GAN generator to clearly indicate important segments of the video. The Actor learns the key-shot selection strategy, and the Critic learns a value function to score the Actor's selection.

2) Weakly-supervised learning: In order to alleviate the need for extensive human-generated ground-truth data, weakly-supervised learning uses less-expensive weak annotation such as video-level or shot-level annotated data to train VS models.

Obtaining video-level annotations from web videos is a common method for weakly-supervised VS, since these web videos can be easily obtained from various video platforms (e.g., YouTube), and the video-level annotations can be collected using a set of topic labels as search keywords. As early as 2017, Panda *et al.* (2017) [110] propose a weakly-supervised VS framework to train models through video-level annotation. They first use multiple web videos of the same category to train a parametric model to classify a given video (because similar videos often have similar summaries) and then generate an importance map based on the video sequence and its category. Li *et al.* (2021) [111]'s methods is similar to that of Panda *et al.* (2017) [110]. They notice that the video classification sub-network can also generate video representations, so they use the video classification sub-network to further narrow the distance between the summary video and the original video representation. Cai *et al.* (2018) [112] use two encoders to generate attention scores of the original video and hidden variables of the web video. Then the decoder generates a summary based on the attention scores and the hidden variables.

In addition to making use of a large number of web videos, some other methods are used for weakly-supervised VS. Ho *et al.* (2018) [113] make an attempt in the field of cross-domain VS. They transfer the model trained with the annotated third-person videos to the first-person video domain with only a small amount of annotated data. The specific method is to use a shared encoder to extract the domain-invariant information, use a unique encoder in each domain to extract the domain-specific information, and then reduce the semantic difference between the two domains through a difference loss. Chen *et al.* (2019) [114] divide a single video into N sequences of equal length, and each sequence is processed at two levels in turn. The network of each level consists of an LSTM, where the first is trained using sequence-level annotations, and the second predicts importance scores based on previously given scores. In addition, they define a sub-reward function when processing each subsequence to solve the sparse reward problem in the global reward function in previous methods. BackTAL (2021) [115] learns instance-level action patterns from video-level labels. To overcome the confusion between actions and backgrounds, it uses a separation loss to guide the separation of action responses from background responses, effectively learning an action localization model that reduces misjudgments of background frames. ECM (2022) [116] simultaneously learns two mainstream weakly supervised localization-by-classification pipelines: the pre-classification pipeline and the post-classification pipeline. It implements two parallel network within a unified framework to simulate these two pipelines while making the two network share the same classifier.

C. User preference

Considering that different users have different preferences for the same video, and the general VS cannot satisfy the subjectivity of users, Sharghi *et al.* [14], [117] propose a query-focused video summarization, which can generate summaries under the guidance of text queries, and also propose a standard dataset for query-focused video summarization.

1) Textual query-based summarization: Incorporating textual queries into video summarization allows for the creation of summaries that are consistent with the user-specified text inputs, leveraging methods such as attention interaction and language-image pre-training to enhance the relevance of the summary content.

The methods used in the early work are based on Determinantal Point Process (DPP) [118]. DPP is a probabilistic model for selecting subsets from a set, which is widely used in tasks such as document summarization. However, it can only solve general set problems and cannot capture the inherent sequential nature of video tasks. To overcome this shortcoming, Gong *et al.* (2014) [119] propose a sequential

DPP (seqDPP) to model the temporal relationship between video frames in VS, such that whether the previous shot is selected affects whether the next shot is selected. Based on seqDPP, Sharghi *et al.* (2016) [14] propose a Sequential and Hierarchical DPP (SH-DPP), which has two sublayers. The first layer summarizes the shots related to queries, and the second layer summarizes the shots unrelated to queries. Furthermore, Sharghi *et al.* (2017) [117] propose a Query Conditioned Sequential DPP (QC-DPP), which uses a memory network to parameterize the DPP kernel. In this way, the user's query focuses on different frames of each shot, achieving better results.

In addition to the DDP-based methods, Zhang *et al.* (2018) [120] apply generative adversarial networks to facilitate query-focused video summarization. The generator models the temporal dependency between shots through a Bidirectional LSTM [53]. The discriminator distinguishes the ground-truth summary from the random and generated summary. The role of the random summary is to prevent the generator from generating random trivial short sequences.

With the outstanding performance of attention mechanism [50] in various visual tasks, researchers begin to apply it to query-focused video summarization. Jiang *et al.* (2019) [121] first apply the attention mechanism to query-focused video summarization. After extracting shot-level features, they use the attention mechanism to determine which shots in the scene are most relevant to the query. Xiao *et al.* (2020) [122], [123] also use the attention mechanism to integrate the semantic information of the query. Taking the Convolutional Hierarchical Attention Network (CHAN) as an example, they first calculate the local attention within each shot and the global attention among different shots and then aggregate the local attentive features and the global attentive features. According to the distance similarity between the aggregated features and the query vector, the correlation score is obtained. In addition, they improve the calculation efficiency in two ways: using 1D convolution to reduce the dimension of shot-level features, and sharing the same weight matrix when calculating the attention in different shots.

The network structure of the Query-Biased Self-Attentive Network (QSAN) [122] is similar to that of CHAN. They select critical frames and query-related frames successively. Specifically, they first train the model using a large number of video caption datasets to guide the model to obtain a more representative summary, and then transfer the trained model to the query-focused video summarization, add query information, and generate video summaries related to the query. However, the video caption can only guide the model to generate a generic summary. It does not integrate the query information, resulting in the final effect being inferior to CHAN, because the frames that are considered unnecessary in the generic summary may also be related to the query.

Further, Yuan *et al.* (2019) [124] establish an attention interaction between the sentence query and video sequence, and then perform GCN to model sentence-specific relationships among different video clips. Huang *et al.* (2023) [125] utilize a semantic booster built on the Transformer architecture to mitigate the ineffective semantic embedding in textual queries, and map frame-level human expert scores to segment-level pseudo scores.

Otherwise, Nalla *et al.* (2020) [126] use Inflated 3D ConvNet (I3D) [127] to extract shot-level features to capture the long spatial and temporal dependency. Huang *et al.* (2023) [128] considers causal relationships and models the behavior of key components in the video summarization problem from four aspects, improving the interpretability of the model. Li *et al.* (2023) [129] introduces multi-modal self-supervised learning for pre-training and proposes a video summarization method that progressively emphasizes important content in the input video sequence. Krubiński *et al.* (2023) [130] propose a new multimodal article summarization dataset and transfer knowledge from simpler text-to-text summarization tasks as pre-training.

Some researchers have also attempted to explore the potential of CLIP [131] in query-focused video summarization, due to its excellent alignment of visual and textual modalities. Narasimhan *et al.* (2021) [132] propose CLIP-It that utilizes the image and text encoders of the language-image pre-training model CLIP [131] to encode video frames and queries. Then it fuses the features of the two and gives importance scores through the attention mechanism, achieving state-of-the-art on the query-focused video summarization dataset. In addition, CLIP-It uses the description generated by the video description generator as text input to obtain a generic video summary when there are no user queries. Han *et al.* (2024) [133] fine-tunes CLIP to leverage its multimodal ability and proposes saliency pooling which simply average-pooling saliency scores from neighboring frames.

More recently, Yang *et al.* (2024) [134] introduce a task-decoupled unit to capture task-specific and shared representations and proposes an inter-task feedback mechanism to utilize the interplay between two tasks. Jiang *et al.* (2024) [135] propose a transformer-based network to utilize multiple input modalities and employ a contrastive objective to enclose the modality gap. Zhou *et al.* (2024) [136] introduce subtask prior features to constrain joint task features, thus balancing the inter-task correlations, complementarity, and differences in joint tasks. Xiao *et al.* (2024) [137] attempts to perform multi-scale interactions within and between modalities, achieving local perception of multimodal and temporal relationships, as well as the accumulation of global video knowledge. V2Xum-LLaMA [138] employs the CLIP visual encoder to encode video features, integrates text tokens with visual tokens, and utilizes LLaMA-2 [139] as the text decoder to generate video summaries, text summaries, or a combination of both, based on temporal prompts and task prompts.

2) Visual query-based summarization: Generating video summaries based on visual queries focuses on using the visual information provided by users to guide the summarization process, ensuring that the generated summaries are closely aligned with the user's visual preferences and expectations.

Wu *et al.* (2022) [140] constructs query and visual features into an ego graph and represents the query as a more fine-grained representation through edge convolution [141], which is named "intent". The user can adjust the value of intent through the user interface to generate a summary more aligned

with the user's expectations. Zhang *et al.* (2023) [142] introduces a cross-modal learning strategy for the first time in video summarization, learning better contextual representations by jointly modeling visual features at both frame and video levels. Sahu *et al.* (2023) [143] transfer the knowledge of auxiliary videos and apply random walks on a constraint graph to extract video summaries. Fei *et al.* (2023) [144] includes MI-based shot-boundary detection and semantical event detection to generate personalized video summaries according to web videos or images.

Multi-video summarization aims to extract a summary that represents the common or key content from multiple videos, where each video can be regarded as a query relative to the others. Wu et al. (2020) [145] present a dynamic graph convolutional network (mvsDGCN) for addressing the multi-video summarization task. Unlike previous methods using fixed graph Laplacian matrices, mvsDGCN updates the adjacency matrix at each layer to better capture the dynamic dependencies of graph nodes. Messaoud et al. (2021) [146] propose a hierarchical attention mechanism for video-query cross-modal modeling and pointer networks to avoid frame reordering, jointly optimizing conciseness, temporal coherence, and query relevance. QDAVOL (2023) [147] utilizes tag information and web images to identify query intent, and employs event-based object detection and grouping along with the African vulture optimization algorithm (AVOA) for efficient key frame selection. Barbieri *et al.* (2021) [148]introduced human-strategy-based content selection criteria for video summarization, encompassing inclusion (introduction, chronology, fluency, redundancy, uniqueness, coverage, visual support, sources) and exclusion criteria (subjectivity, supplementary content, commentary, interviews) to improve multi-video summary relevance and user satisfaction.

D. High Precision

Some video summarization techniques focus on the accuracy of extracting key video content, and their applications generally do not rely on user subjectivity or the excitement level of the content. For example, videos in the medical field require precise localization of abnormal moments, while videos in the sports domain focus on shots and close-ups of goals or scores.

1) Medical video summarization: Medical video summarization aims to extract critical clinical information from medical videos, and often employs methods such as reinforcement learning to generate summaries that are helpful for diagnosis and monitoring.

Sabha *et al.* (2023) [149] use a pre-trained DCNN to classify video frames and create video summaries with and without face masks for monitoring COVID-19 face mask protocols in crowded places. Liu *et al.* (2020) [150] propose a reinforcement learning-based ultrasound video summarization framework with a new diagnostic view plane reward to encourage agents to select more necessary clinical information. Zhu *et al.* (2001) [151] propose a hierarchical medical video summarization strategy, which combines clustering, group correlation, and semantic analysis to generate summaries with content and granularity that vary from person to person. Mathews *et al.* (2022) [152] propose an unsupervised reinforcement learning video summarization framework that encodes ultrasound video frames from three aspects: an anomaly classification encoder, a segmentation encoder, and a convolutional autoencoder, and uses structural similarity index matrix and health-based classification as rewards.

2) Sports video summarization: Sports video summarization focuses on identifying and highlighting key moments in sports footage by detecting excitement scores, segmenting dynamic plays, and extracting replay segments to provide concise and engaging highlights of the games.

Raval *et al.* (2023) [153] propose to detect cricket video shot boundaries and extract replay segments from the cricket video sequence. Narwal *et al.* (2023) [154] propose a dual neural network for dynamic video segmentation and key segment detection, relying on audio analysis in cricket videos. Banjar *et al.* (2024) [155] proposes to extract keyframes in sports videos by detecting the excitement score in the audio stream. Davids *et al.* (2024) [156] proposes a hybrid video summarization framework to detect and label exciting events, and eliminates replay shots to generate a concise summary. Bhat *et al.* (2023) [157] generate high-quality cricket video highlights, textual summary, and audio summary from the original video in any of these three formats.

IV. Evaluation

A. Dataset

In existing studies of VS, various video datasets have been proposed for different downstream tasks. Taking into account both video quality and citation, we summarize the commonly used video summarization datasets, as shown in Table II. Among them, SumMe, TVSum, and YouTube are the most commonly used benchmark datasets, and are divided into three settings: canonical, augmented, and transfer, to comprehensively evaluate the performance and generalization ability of the model, as shown in Table III.

The Query-focused Video Summarization dataset (QFVS) includes videos and queries. The videos are selected from the UT Egocentric (UTE) dataset [163], including four videos captured by head-mounted cameras in a natural and uncontrolled environment at 15 fps with 320×480 resolution. The average duration of each video is about 35 hours, covering various activities such as eating, shopping, attending lectures, driving, and cooking.

To construct the queries for QFVS, Sharghi *et al.* (2017) [117] first build a lexicon containing 48 common concepts, with two concepts paired together to form a query, making a total of 46 queries. The videos are uniformly segmented into 5-second clips, from which 5 frames are extracted. If the concept appears in any of these frames, it is considered relevant to the clip. Taking into account the subjectivity of annotators, three annotators are asked to annotate separately and the union of all the annotations is taken as the final annotation of each shot. For the four UTE videos, they finally obtain 4.13, 3.95, 3.18, and 3.62 concepts per shot on average.

Ji*et al.* (2019) [161] select ten hot search terms from Wikipedia as text queries and collect 100 videos for each

TABLE II
MAIN CHARACTERISTICS OF VIDEO SUMMARIZATION DATASETS.

Dataset	Size	Duration	Annotators	Camera View	Content Themes
OVP [158]	50	1-4 min	5	Third-person view	documentary, educational, ephemeral, historical, lecture
Youtube [158]	50	1-10 min	5	Third-person view	cartoons, news, sports, commercials, tv-shows, home videos
SumMe [159]	25	1-6 min	15-18	Third-person view	holidays, events, sports
TVSum [106]	50	2-10 min	20	Third-person view	news, how-to's, documentaries, user-generated content (vlog, egocentric)
CoSum [160]	51	$\leq$4 min	3	Third-person view	holidays, events, sports (10 categories from SumMe)
MED [48]	160	1-5 min	1-4	Third-person view	holiday, parade, celebration, cooking, work and other 15 genres
FPVSum [113]	98	4.85 min	13-25	First-person view	bike, bikepolo, boxing, horse, umping, longboarding, motor, parkour, plane, rock climbing, scuba, skate, ski, surfing
QFVS [117]	4	3-5 hours	138	First-person view	eating, shopping, attending lectures, driving, cooking
MVS1K [161]	1000	$\leq$4 min	4	Third-person	10 hot events from Wikipedia News from the year of 2011 to 2016
QCVS [162]	190	2-3 min	$\geq$5	Third-person	the top YouTube queries between 2008 and 2016 from 22 different categories

search term. Following the same rules as Vasudevan et al. [164], Huang*et al.* (2020) [162] choose 22 hot search terms from YouTube, use YouTube's auto-complete function to obtain more realistic and longer corresponding text queries, and collect the top video results with a duration of 2 to 3 minutes.

B. Metric

Reference-Based Evaluation: The evaluation protocol used in most current studies compares the generated summary with the ground-truth summary provided by the dataset, that is, to compute the F_β-measure between the generated summary and the ground-truth summary. Given a generated summary $S = \{s_1, ..., s_N\}$ and a ground-truth summary $G = \{g_1, ..., g_N\}$, where $s_i \in (0, 1), g_i \in (0, 1)$. The precision $P = \sum_{n=1}^{N} s_i \cdot g_i / \sum_{n=1}^{N} s_i$, recall $P = \sum_{n=1}^{N} s_i \cdot g_i / \sum_{n=1}^{N} g_i$, and F_β-measure are calculated by

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}, \tag{3}$$

where β balances the relative importance between P and R, and is commonly set to $\beta = 1$. So F_1-measure is the harmonic mean of P and R.

Furthermore, different annotators typically provide multiple ground-truth summaries for the same video. In SumMe [99], the maximum F_1-measure between the generated summary and multiple ground-truth summaries is taken for evaluation. In TVSum [106], the average F_1-measure between the generated summary and multiple ground-truth summaries is taken for evaluation. Finally, the average F_1-measure of all videos in the dataset is taken as the performance of the method on a particular dataset.

TABLE III
DIFFERENT SETTINGS ON THE SUMME AND TVSUM DATASETS.

Dataset	Settings	Training & Validation	Testing
SumMe	Canonical	80 % SumMe	20 % SumMe
	Augmented	OVP + Youtube + TVSum + 80 % SumMe	20 % SumMe
	Transfer	OVP + Youtube + TVSum	SumMe
TVSum	Canonical	80 % TVSum	20 % TVSum
	Augmented	OVP + Youtube + SumMe + 80 % TVSum	20 % TVSum
	Transfer	OVP + Youtube + SumMe	TVSum

Rank-Based Evaluation: Otani *et al.* (2019) [165] point out that the widely used F_1-measure can be interfered by the video segmentation method. Under a specific segmentation method, even randomly generated summaries can achieve a high F_1-measure. To more objectively explore the correlation between the ground-truth and generated summaries, the rank correlation coefficients are presented. Kendall correlation coefficient τ [166] and Spearman correlation coefficient ρ [167] can measure the similarities between the rankings, where the former is based on the relationship between sample pairs, and the latter is based on the rank difference.

Given a video with N frames, its predicted importance score sequence generated by the summarization model is presented by $X = [x_1, x_2, ..., x_N]$, and its ground-truth importance score sequence is presented by $Y = [y_1, y_2, ..., y_N]$. The Kendall correlation coefficient τ is given by

$$\tau = \frac{C - D}{\sqrt{(C + D + T_x) \times (C + D + T_y)}}, \tag{4}$$

where C denotes the number of consistent data pairs, D denotes the number of divergent data pairs, T_x denotes the number of parallel permutations in sequence X, and T_y denotes the number of parallel permutations in sequence Y. Note that if parallel permutations occur in both X and Y, neither T_x nor T_y can be counted. Assume that the elements at the same position in the two sequences are regarded as a data pair, denoted as (x_i, y_i).

Following the same definition, the Spearman correlation

coefficient ρ can be formulated as

$$\rho = \frac{\frac{1}{N}\sum_{i=1}^{N}(R[x_i] - \overline{R[x]})(R[y_i] - \overline{R[y]})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(R[x_i] - \overline{R[x]})^2 \times \frac{1}{N}\sum_{i=1}^{N}(R[y_i] - \overline{R[y]})^2}}, \quad (5)$$

where $R[x_i]$ denotes the rank of X_i in X, and $\overline{R[x]}$ denotes the average rank of X. Similarly, $R[y_i]$ and $\overline{R[y]}$ can be obtained.

To verify the effectiveness of the correlation coefficient, two types of experiments are typically conducted: random and human settings. In the random setting, the correlation coefficient is close to 0 because it compares a randomly generated importance score sequence with one given by an annotator, resulting in almost no correlation. In the human setting, the correlation coefficient is highest because both importance score sequences are manually annotated, showing a strong correlation.

C. Comparison

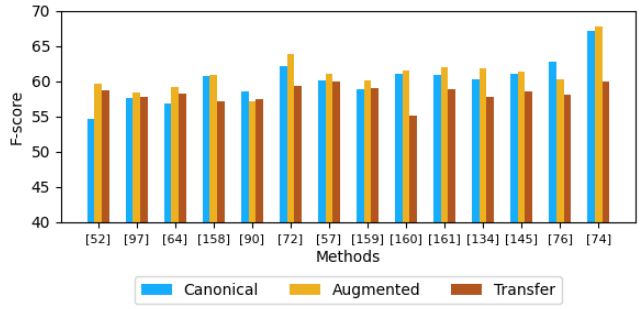


Fig. 5. Performance of techiniques on TVSum dataset.

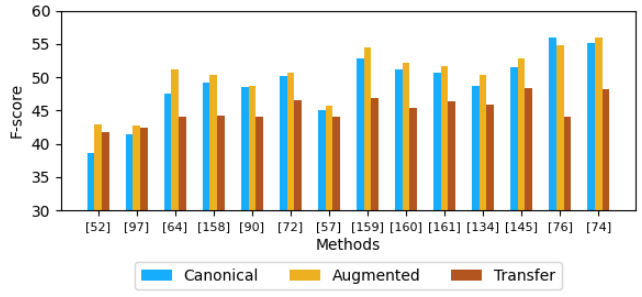


Fig. 6. Performance of techiniques on SumMe dataset.

In this section, we present the performance of representative methods selected from section 4. The results of Figure 5, Figure 6, Table IV and Table V are cited from the original papers. Figure 5 and 6 illustrate the F_1-measure on the TVSum and SumMe datasets, respectively. The results are cited from the original papers. From the results, we can make the following observations.

The effectiveness of attention mechanisms in video summarization is evident, as methods utilizing attention (SUM-GDA [168], LMHA [169], VJMHT [170], SA-CFT [76], STVT [74]) consistently outperform traditional approaches. Compared to LSTM-based methods (vsLSTM [52], dppLSTM [52]), GCN-based methods (RSGN [57]), and FCN-based methods (SUM-FCN [64]), attention-based models excel in capturing long-range dependencies and dynamically weighting informative frames, leading to more accurate importance score predictions. This is further supported by qualitative results on TVSum (Figure 7), where attention-based methods (VASNet [71], DSNet [72], DR-DSN [98]) generate importance scores that align more closely with ground-truth annotations than GAN-based methods (SUM-GAN, SUM-GAN-AAE [94]). The latter often suffer from instability in adversarial training, which can lead to inconsistent or noisy predictions. Methods leveraging spatio-temporal modeling, such as 3D convolution or graph convolution, also achieve high performance by explicitly capturing temporal dynamics and frame relationships. However, their computational cost is significantly higher than that of 2D convolution-based approaches, limiting their scalability for real-world applications. This trade-off between performance and efficiency suggests a need for lightweight spatio-temporal modeling techniques in future work.

A key observation is that models trained under canonical settings (same-domain training and testing) generally outperform those under transfer settings (cross-domain evaluation), as seen in models trained on YouTube and OVP datasets that exhibit poor generalization to TVSum and SumMe due to domain gaps in video summarization. This gap arises from dataset bias, where variations in video content, editing styles, and annotation criteria cause distribution shifts, compounded by limited training data that leads to overfitting, particularly in complex architectures like GANs or graph-based models. Additionally, the subjectivity of human-annotated importance scores introduces noise and inconsistency across datasets, further hindering cross-domain adaptation. The performance gap between canonical and enhanced settings (e.g., augmented or fine-tuned models) highlights the need for data diversity and domain adaptation techniques, suggesting future research explore semi-supervised learning, domain-invariant feature learning, or synthetic data generation to address these challenges.

TABLE IV
COMPARISON (RANK CORRELATION COEFFICIENTS) OF VIDEO SUMMARIZATION METHODS.

Method	Kendall's τ		Spearman's ρ	
	SumMe	TVSum	SumMe	TVSum
Random	0.000	0.000	0.000	0.000
DR-DSN [98]	0.047	0.020	0.048	0.026
CSNet [91]	-	0.025	-	0.034
dppLSTM [52]	-	0.042	-	0.055
$RSGN_{uns}$ [57]	0.071	0.048	0.073	0.052
RSGN [57]	0.083	0.083	0.085	0.090
CSNet+GL+RPE [107]	-	0.070	-	0.091
Hierarchical RL [114]	-	0.078	-	0.116
CLIP-It [132]	-	0.108	-	0.147
Human	0.205	0.177	0.213	0.204

It is noteworthy that some methods may exhibit significant performance differences on different evaluation metrics. The results in Table IV show the rank correlation coefficients

achieved by various methods on these two datasets. For example, DR-DSN achieves a higher F_1-measure on TVSum than dppLSTM, but its correlation coefficient is lower than that of dppLSTM. Taking into account the calculation of the two evaluation metrics, the F_1-measure focuses on whether the selected frames exist in the set of ground-truth frames, while the rank correlation coefficient concerns whether the importance score curve of the video frames aligns with the true importance curve. We believe that the reason for this phenomenon is that some video frames have higher importance scores, but the gradient of the scores does not conform to the ground-truth importance scores. To ensure more accurate evaluation, future work should utilize both of these metrics concurrently to assess the performance of video summarization methods.

Table V shows the performance of existing query-focused video summarization methods. In query-centric video summarization methods, attention mechanisms are the most commonly used approach for integrating visual and semantic information. We observe that integrating video features with query features at earlier stages is more likely to achieve higher F_1-measure. Both QSAN [122] and CHAN [123] integrate query vectors during the calculation of attention between shots. In contrast, CLIP-It incorporates query vectors while calculating the attention within each shot, resulting in an improved F_1-measure. The high performance of CLIP-It is also attributed to the image-text similarity alignment capability within the pre-trained CLIP.

TABLE V
F-MEASURE(%) ON QFVS DATASET.

Method	Video 1	Video 2	Video 3	Video 4	AVG
seqDPP [119]	36.59	43.67	25.26	18.15	30.92
SH-DPP [14]	35.67	42.72	36.51	18.62	33.38
QC-DPP [117]	48.68	41.66	56.47	29.96	44.19
TPAN [120]	48.74	45.30	56.51	33.64	46.05
QSAN [122]	48.52	46.64	56.93	34.25	46.59
CHAN [123]	49.14	46.53	58.65	33.42	46.94
HVN [121]	51.45	47.49	61.08	35.47	48.87
SVUI [126]	50.96	48.28	58.41	39.18	49.20
IntentVizor [140]	51.27	53.48	61.58	37.25	50.90
CLIP-It [132]	**57.13**	**53.60**	**66.08**	**41.41**	**54.55**

V. FUTURE OUTLOOK

A. Interactivity

Interactivity is one of the essential characters that video summarization needs to possess. Most existing video summarization methods lack interaction with the user. Once a video summary is generated, users cannot modify it. Query-based video summarization focuses on the content that users want from the video, but the queries are often combinations of words or sentences constructed using prompt templates, which limits the flexibility of interaction to some extent.

The goal of enhancing interactivity is to obtain customized summaries that meet the diverse preferences of users. On one hand, the modality of user queries can be expanded, extending beyond text and image modalities [140], [143] to include more diverse forms. On the other hand, previous query-based video summarization has focused on the thematic content while neglecting other elements of the video. For example, in movie scenes, aspects such as special camera movements, important characters, and key dialogues are deeper semantic information that has not been fully explored by existing works.

Moreover, controllable video summarization is another approach to enhance interactivity. Such methods would allow for user control signals after the summary is generated. The initial summary could be adjusted based on user feedback until the current summary sufficiently meets the user's desired output.

B. Generalization

Video summarization methods exhibit significant differences across videos from various domains. Generalizing models trained on a source domain to a new target domain is a critical challenge in video summarization. Despite the limited datasets available for video summarization, there are extensive video resources across different domains for other video tasks.

Transfer learning and incremental learning can effectively balance the relationship between new and old knowledge, addressing domain gaps. For instance, using importance propagation algorithms to extend video localization data to video summarization tasks [171]. Coordinating more video tasks with video summarization to promote the performance of video understanding models in various downstream tasks is a feasible future direction.

Additionally, leveraging existing knowledge from pre-trained models for downstream task inference is another way to improve generalization capabilities. When the training data for pre-trained models is sufficient, performing few-shot or zero-shot inference on new domains can yield significant results. For instance, pre-training vision-language model can enhance the generalization capability of query-focused video summarization across different video domains and expand the scope of query texts [132].

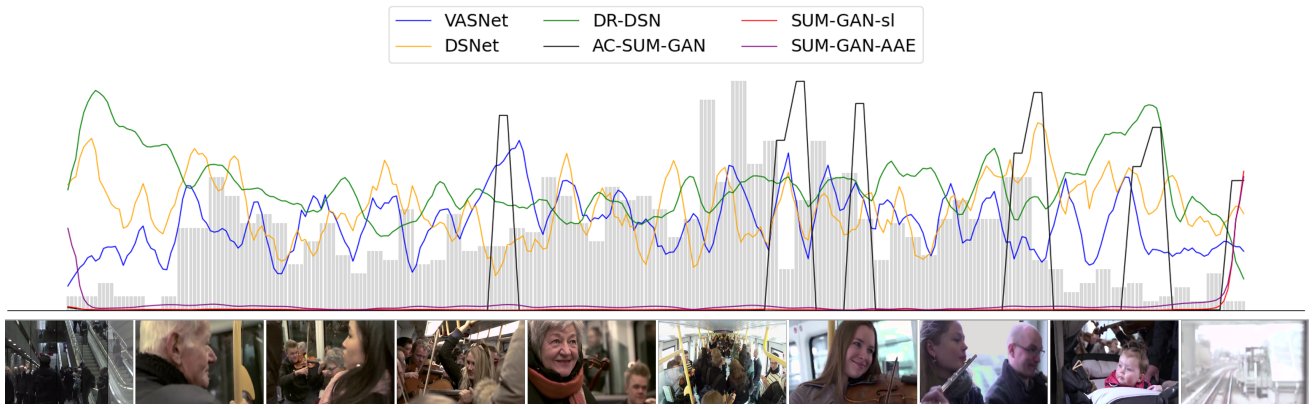
C. Interpretability

Interpretability has always been a problem faced by deep learning based models. With the widespread application of deep learning methods in video summarization, interpretability has become an important issue to be solved. The problem is that during the inference process of the video summary model, we cannot know the meaning of the model's output features, nor can we analyze the reasons for the different importance of video frames.

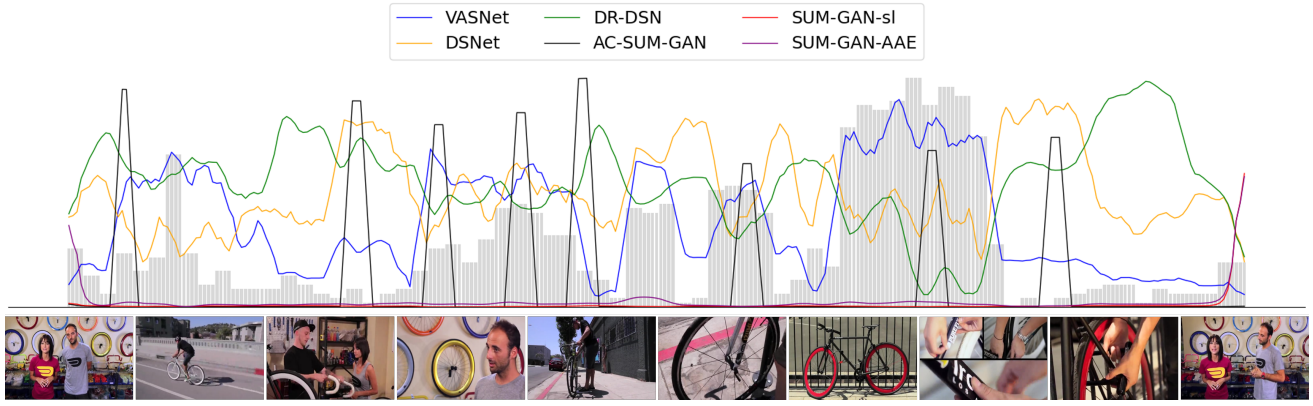
The most direct way to improve the interpretability of the model is to visualize and analyze internal representations at different stages. It is possible to evaluate the causal effect of different internal representations on the results through causal reasoning and explore the meaning of the "importance" understood by the model in real videos.

Interpretability has always been a challenge for deep learning-based models. With the widespread application of deep learning methods in video summarization, interpretability has become an important issue that needs to be addressed.

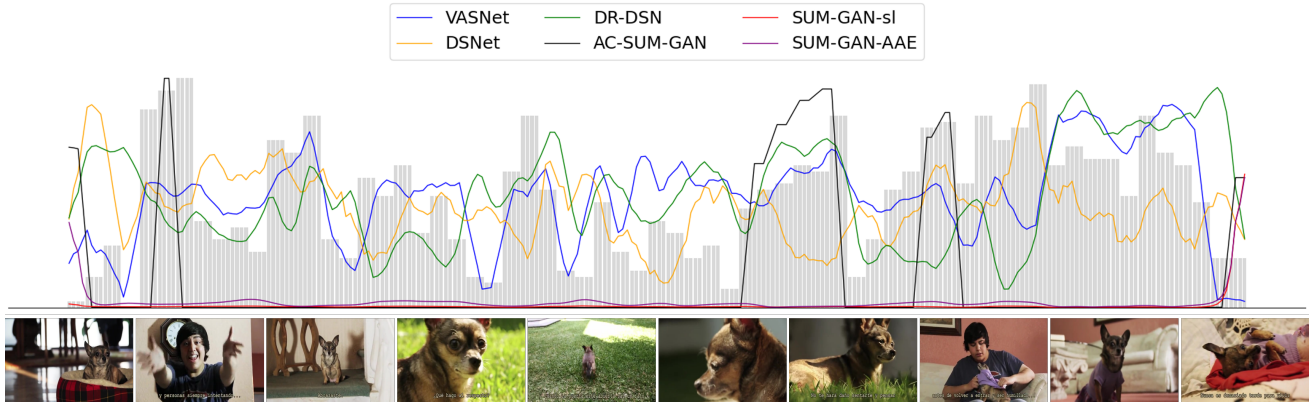
Video summarization methods help in identifying the most critical parts and removing redundant sections, but they do not



(a) Video #35 in TVSum.



(b) Video #44 in TVSum.



(c) Video #48 in TVSum.

Fig. 7. Qualitative results of different video summarization methods.

provide reasons why a particular video segment is considered important or redundant. Consequently, the resulting summarized videos may lack sufficient persuasiveness. The most direct way to improve the explainability of models is through the visualization and analysis of internal representations at different stages. A more persuasive approach would involve providing reasons for the importance of significant visual content and explaining why certain visual content is deemed unimportant based on its failure to meet specific criteria [172].

Causal inference can evaluate the causal impact of different internal representations on the outcome. Using memory mechanisms can aid in understanding the causal relationships of sequential events in a video [173], but this has not yet received much attention in the field of video summarization.

D. Multimodal Fusion

A complete video encompasses multiple forms of information (audio, speech, text, and images). Viewers typically browse the entire original video from start to finish and annotate it based on its comprehensive content. However, existing methods often consider only the visual information,

neglecting the role of other modalities in determining key content.

Query-focused video summarization can use textual information as query conditions, but the queries are usually composed of single-form phrases or template sentences. Audio and speech information also contribute to identifying key shots in the video. Utilizing various feature extractors to process multi-modal information, and finding the spatio-temporal correspondences between different modalities through modality alignment and fusion, can help distinguish important from unimportant video shots. For example, voice clustering and speech recognition can help retrieve specific character segments, even if the character does not appear on screen. Additionally, the output of video summarization is a subset of the original video frames. Using synthesized voice narration instead of the original soundtrack to describe the generated video summary can provide a better user experience. Therefore, integrating multi-modal information at both the input and output ends is a feasible method to improve the performance and user experience of video summarization methods.

E. Long-Form Video

Long-form video summarization, particularly for complex content such as movies, faces multifaceted challenges. Firstly, long videos like movies typically contain multiple intertwined storylines and characters, requiring models to accurately capture and integrate these key elements. Secondly, the large temporal span leads to uneven information distribution within the same video. Additionally, the long-term dependencies require models to have robust long-term memory capabilities. Finally, the specific cultural backgrounds and deep contextual hierarchy on videos require models to have cross-cultural comprehension abilities to accurately capture metaphors, symbols, and other profound meanings.

Representing videos as structured unified memories through text descriptions and features, and leveraging the capabilities of large language models for interactive reasoning and question-answering [173], aids in comprehending complex events and temporal dependencies within long videos. Based on video understanding, extracting key events and establishing causal relationships between them helps capture the main threads of the complete video within the constraints of limited visual information.

VI. Conclusion

We have provided a systematic review of video summarization, covering challenges, methodologies, benchmark datasets, evaluation metrics, and future outlook. We categorize existing methods based on the most critical issues in the field, namely data scarcity, temporal dependency, user preference, and high precision. Furthermore, we have analyzed representative methods through both qualitative and quantitative experiments. Last but not least, we have discussed the remaining challenges and potential future research directions in video summarization.

References

[1] M. V. Mussel Cirne and H. Pedrini, "Viscom: A robust video summarization approach using color co-occurrence matrices," *Multimedia Tools and Applications*, vol. 77, pp. 857–875, 2018.

[2] A. Emad, F. Bassel, M. Refaat, M. Abdelhamed, N. Shorim, and A. AbdelRaouf, "Automatic video summarization with timestamps using natural language processing text fusion," in *2021 IEEE 11th annual computing and communication workshop and conference (CCWC)*. IEEE, 2021, pp. 0060–0066.

[3] R. Panda and A. K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2010–2021, 2017.

[4] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.

[5] H. Li, J. Zhu, J. Zhang, X. He, and C. Zong, "Multimodal sentence summarization via multimodal selective encoding," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5655–5667.

[6] L. Jing, Y. Li, J. Xu, Y. Yu, P. Shen, and X. Song, "Vision enhanced generative pre-trained language model for multimodal sentence summarization," *Machine Intelligence Research*, vol. 20, no. 2, pp. 289–298, 2023.

[7] D. Lin, L. Jing, X. Song, M. Liu, T. Sun, and L. Nie, "Adapting generative pretrained language model for open-domain multimodal sentence summarization," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 195–204.

[8] T. Yu, W. Dai, Z. Liu, and P. Fung, "Vision guided generative pre-trained language models for multimodal abstractive summarization," *arXiv preprint arXiv:2109.02401*, 2021.

[9] Y. Gong, X. Luo, K. Q. Zhu, W. Ou, Z. Li, and L. Duan, "Automatic generation of chinese short product titles for mobile display," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9460–9465.

[10] X. Song, L. Jing, D. Lin, Z. Zhao, H. Chen, and L. Nie, "V2p: vision-to-prompt based multi-modal product summary generation," in *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, 2022, pp. 992–1001.

[11] C. Huang and H. Wang, "A novel key-frames selection framework for comprehensive video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 577–589, 2019.

[12] L. Zhang, L. Sun, W. Wang, and Y. Tian, "Kaas: A standard framework proposal on video skimming," *IEEE Internet Computing*, vol. 20, no. 4, pp. 54–59, 2016.

[13] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou, "Multi-view video summarization," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 717–729, 2010.

[14] A. Sharghi, B. Gong, and M. Shah, "Query-focused extractive video summarization," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 3–19.

[15] J. Almeida, N. J. Leite, and R. d. S. Torres, "Online video summarization on compressed domain," *Journal of Visual Communication and Image Representation*, vol. 24, no. 6, pp. 729–738, 2013.

[16] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and T.-h. Kim, "Event detection based approach for soccer video summarization using machine learning," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 7, no. 2, pp. 63–80, 2012.

[17] R. Agyeman, R. Muhammad, and G. S. Choi, "Soccer video summarization using deep learning," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019, pp. 270–273.

[18] A. Bhalla, A. Ahuja, P. Pant, and A. Mittal, "A multimodal approach for automatic cricket video summarization," in *2019 6th international conference on signal processing and integrated networks (SPIN)*. IEEE, 2019, pp. 146–150.

[19] H. Shingrakhia and H. Patel, "Sgrnn-am and hrf-dbn: a hybrid machine learning model for cricket video summarization," *The Visual Computer*, vol. 38, no. 7, pp. 2285–2301, 2022.

[20] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury, "Context-aware surveillance video summarization," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5469–5478, 2016.

[21] O. Elharrouss, N. Al-Maadeed, and S. Al-Maadeed, "Video summarization based on motion detection for surveillance systems," in *2019*

15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019, pp. 366–371.

[22] K. Muhammad, T. Hussain, and S. W. Baik, "Efficient cnn based summarization of surveillance videos for resource-constrained devices," *Pattern Recognition Letters*, vol. 130, pp. 370–375, 2020.

[23] M. I. Leszczuk and M. Duplaga, "Algorithm for video summarization of bronchoscopy procedures," *Biomedical engineering online*, vol. 10, pp. 1–17, 2011.

[24] R. Hamza, K. Muhammad, Z. Lv, and F. Titouna, "Secure video summarization framework for personalized wireless capsule endoscopy," *Pervasive and Mobile Computing*, vol. 41, pp. 436–450, 2017.

[25] A. G. Del Molino, C. Tan, J.-H. Lim, and A.-H. Tan, "Summarization of egocentric videos: A comprehensive survey," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 65–76, 2016.

[26] M. Kini and K. Pai, "A survey on video summarization techniques," in *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, vol. 1. IEEE, 2019, pp. 1–5.

[27] D. Sen, B. Raman *et al.*, "Video skimming: Taxonomy and comprehensive survey," *arXiv preprint arXiv:1909.12948*, 2019.

[28] M. Basavarajaiah and P. Sharma, "Survey of compressed domain video summarization techniques," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–29, 2019.

[29] H. B. U. Haq, M. Asif, and M. B. Ahmad, "Video summarization techniques: a review," *Int. J. Sci. Technol. Res*, vol. 9, no. 11, pp. 146–153, 2020.

[30] V. Raut and R. Gunjan, "Video summarization approaches in wireless capsule endoscopy: a review," in *E3S web of conferences*, vol. 170. EDP Sciences, 2020, p. 03005.

[31] T. Hussain, K. Muhammad, W. Ding, J. Lloret, S. W. Baik, and V. H. C. De Albuquerque, "A comprehensive survey of multi-view video summarization," *Pattern Recognition*, vol. 109, p. 107567, 2021.

[32] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1838–1863, 2021.

[33] V. Vasudevan and M. Sellappa Gounder, "Advances in sports video summarization–a review based on cricket videos," in *Advances and Trends in Artificial Intelligence. From Theory to Practice: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part II 34*. Springer, 2021, pp. 347–359.

[34] V. Tiwari and C. Bhatnagar, "A survey of recent work on video summarization: approaches and techniques," *Multimedia Tools and Applications*, vol. 80, no. 18, pp. 27 187–27 221, 2021.

[35] A. S. Parihar, R. Mittal, P. Jain *et al.*, "Survey and comparison of video summarization techniques," in *2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)*. IEEE, 2021, pp. 268–272.

[36] P. Narwal, N. Duhan, and K. K. Bhatia, "A comprehensive survey and mathematical insights towards video summarization," *Journal of Visual Communication and Image Representation*, vol. 89, p. 103670, 2022.

[37] K. R. Raval and M. M. Goyani, "A survey on event detection based video summarization for cricket," *Multimedia Tools and Applications*, vol. 81, no. 20, pp. 29 253–29 281, 2022.

[38] H. Moussaoui, N. El Akkad, and M. Benslimane, "A review of video summarization," in *International Conference on Digital Technologies and Applications*. Springer, 2023, pp. 516–525.

[39] T. Correia, A. Cunha, and P. Coelho, "A review on the video summarization and glaucoma detection," in *International Conference on Wireless Mobile Communication and Healthcare*. Springer, 2023, pp. 144–156.

[40] P. Meena, H. Kumar, and S. K. Yadav, "A review on video summarization techniques," *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105667, 2023.

[41] A. Sabha and A. Selwal, "Data-driven enabled approaches for criteria-based video summarization: a comprehensive survey, taxonomy, and future directions," *Multimedia Tools and Applications*, vol. 82, no. 21, pp. 32 635–32 709, 2023.

[42] P. G. Shambharkar and R. Goel, "From video summarization to real time video summarization in smart cities and beyond: A survey," *Frontiers in big Data*, vol. 5, p. 1106776, 2023.

[43] M. Peronikolis and C. Panagiotakis, "Personalized video summarization: A comprehensive survey of methods and datasets," *Applied Sciences*, vol. 14, no. 11, p. 4400, 2024.

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

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[48] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*. Springer, 2014, pp. 540–555.

[49] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986," *Biometrika*, vol. 71, no. 599-607, p. 6, 1986.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[51] R. Merkle and M. Hellman, "Hiding information and signatures in trapdoor knapsacks," *IEEE transactions on Information Theory*, vol. 24, no. 5, pp. 525–530, 1978.

[52] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*. Springer, 2016, pp. 766–782.

[53] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm networks," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 4. IEEE, 2005, pp. 2047–2052.

[54] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[55] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 863–871.

[56] S. Huang, X. Li, Z. Zhang, F. Wu, and J. Han, "User-ranking video summarization with multi-stage spatio–temporal representation," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2654–2664, 2018.

[57] B. Zhao, H. Li, X. Lu, and X. Li, "Reconstructive sequence-graph network for video summarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2793–2801, 2021.

[58] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[59] W. Zhu, Y. Han, J. Lu, and J. Zhou, "Relational reasoning over spatial-temporal graphs for video summarization," *IEEE Transactions on Image Processing*, vol. 31, pp. 3017–3031, 2022.

[60] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[61] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder–decoder networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1709–1717, 2019.

[62] H. Khan, T. Hussain, S. U. Khan, Z. A. Khan, and S. W. Baik, "Deep multi-scale pyramidal features network for supervised video summarization," *Expert Systems with Applications*, vol. 237, p. 121288, 2024.

[63] Y. Jin, X. Tian, Z. Zhang, P. Liu, and X. Tang, "C2f: An effective coarse-to-fine network for video summarization," *Image and Vision Computing*, p. 104962, 2024.

[64] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 347–363.

[65] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[66] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, and B. Kainz, "Video summarization through reinforcement learning with a 3d spatio-temporal u-net," *IEEE Transactions on Image Processing*, vol. 31, pp. 1573–1586, 2022.

[67] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *European conference on computer vision*. Springer, 2010, pp. 140–153.

[68] F. Alharbi, S. Habib, W. Albattah, Z. Jan, M. D. Alanazi, and M. Islam, "Effective video summarization using channel attention-assisted encoder–decoder framework," *Symmetry*, vol. 16, no. 6, p. 680, 2024.

[69] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[70] A. Singh and M. Kumar, "Bayesian fuzzy clustering and deep cnn-based automatic video summarization," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 963–1000, 2024.

[71] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 39–54.

[72] W. Zhu, J. Lu, J. Li, and J. Zhou, "Dsnet: A flexible detect-to-summarize network for video summarization," *IEEE Transactions on Image Processing*, vol. 30, pp. 948–962, 2020.

[73] Y. Zhang, Y. Liu, and C. Wu, "Attention-guided multi-granularity fusion model for video summarization," *Expert Systems with Applications*, vol. 249, p. 123568, 2024.

[74] T.-C. Hsu, Y.-S. Liao, and C.-R. Huang, "Video summarization with spatiotemporal vision transformer," *IEEE Transactions on Image Processing*, vol. 32, pp. 3013–3026, 2023.

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

[76] R. Zhong, R. Wang, W. Yao, M. Hu, S. Dong, and A. Munteanu, "Semantic representation and attention alignment for graph information bottleneck in video summarization," *IEEE Transactions on Image Processing*, 2023.

[77] T. Wu, H. Ren, P. Li, and J. Leskovec, "Graph information bottleneck," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 437–20 448, 2020.

[78] X. Song, L. Sun, J. Lei, D. Tao, G. Yuan, and M. Song, "Event-based large scale surveillance video summarization," *Neurocomputing*, vol. 187, pp. 66–74, 2016.

[79] K. Muhammad, T. Hussain, M. Tanveer, G. Sannino, and V. H. C. De Albuquerque, "Cost-effective video summarization using deep cnn with hierarchical weighted fusion for iot surveillance networks," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4455–4463, 2019.

[80] R. Mahum, A. Irtaza, M. Nawaz, T. Nazir, M. Masood, S. Shaikh, and E. A. Nasr, "A robust framework to generate surveillance video summaries using combination of zernike moments and r-transform and deep neural network," *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 13 811–13 835, 2023.

[81] J. Xu, Z. Sun, and C. Ma, "Crowd aware summarization of surveillance videos by deep reinforcement learning," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 6121–6141, 2021.

[82] S. Sharma and D. Goyal, "Enhanced security using video summarization for surveillance system using deep lstm model with k-means clustering technique," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 26, no. 3, pp. 913–925, 2023.

[83] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou, "Timechat: A time-sensitive multimodal large language model for long video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 313–14 323.

[84] P. Papalampidi, S. Koppula, S. Pathak, J. Chiu, J. Heyward, V. Patraucean, J. Shen, A. Miech, A. Zisserman, and A. Nematzdeh, "A simple recipe for contrastively pre-training video-first encoders beyond 16 frames," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 386–14 397.

[85] A. C. Hernandez, M. C. Hernandez, F. G. Ugalde, M. N. Miyatake, and H. P. Meana, "A fast and effective method for static video summarization on compressed domain," *IEEE Latin America Transactions*, vol. 14, no. 11, pp. 4554–4559, 2016.

[86] M. Basavarajaiah and P. Sharma, "Ksumm: a compressed domain technique for video summarization using partial decoding of videos," in *International Conference on Advanced Informatics for Computing Research*. Springer, 2018, pp. 241–252.

[87] M. Fei, W. Jiang, and W. Mao, "A novel compact yet rich key frame creation method for compressed video summarization," *Multimedia Tools and Applications*, vol. 77, pp. 11 957–11 977, 2018.

[88] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 202–211.

[89] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[90] X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan, "Unsupervised video summarization with attentive conditional generative adversarial networks," in *Proceedings of the 27th ACM International Conference on multimedia*, 2019, pp. 2296–2304.

[91] Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon, "Discriminative feature learning for unsupervised video summarization," in *Proceedings of the AAAI Conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8537–8544.

[92] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[93] L. Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, "Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9143–9150.

[94] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Unsupervised video summarization via attention-driven adversarial learning," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part I 26*. Springer, 2020, pp. 492–504.

[95] M. N. Minaidi, C. Papaioannou, and A. Potamianos, "Self-attention based generative adversarial networks for unsupervised video summarization," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 571–575.

[96] Q. Yu, H. Yu, Y. Sun, D. Ding, and M. Jian, "Unsupervised video summarization based on the diffusion model of feature fusion," *IEEE Transactions on Computational Social Systems*, 2024.

[97] M. Abbasi and P. Saeedi, "Adopting self-supervised learning into unsupervised video summarization through restorative score," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 425–429.

[98] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[99] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3090–3098.

[100] J. Lei, Q. Luan, X. Song, X. Liu, D. Tao, and M. Song, "Action parsing-driven video summarization based on reinforcement learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 7, pp. 2126–2137, 2018.

[101] G. Wang, X. Wu, and J. Yan, "Progressive reinforcement learning for video summarization," *Information Sciences*, vol. 655, p. 119888, 2024.

[102] Z. Pang, Y. Nakashima, M. Otani, and H. Nagahara, "Contrastive losses are natural criteria for unsupervised video summarization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2010–2019.

[103] 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.

[104] G. Yaliniz and N. Ikizler-Cinbis, "Using independently recurrent networks for reinforcement learning based unsupervised video summarization," *Multimedia Tools and Applications*, vol. 80, pp. 17 827–17 847, 2021.

[105] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5457–5466.

[106] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5179–5187.

[107] Y. Jung, D. Cho, S. Woo, and I. S. Kweon, "Global-and-local relative position embedding for unsupervised video summarization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV*. Springer, 2020, pp. 167–183.

[108] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[109] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Ac-sum-gan: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Transactions*

on Circuits and Systems for Video Technology, vol. 31, no. 8, pp. 3278–3292, 2020.

[110] R. Panda, A. Das, Z. Wu, J. Ernst, and A. K. Roy-Chowdhury, "Weakly supervised summarization of web videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3657–3666.

[111] Z. Li and 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.

[112] S. Cai, W. Zuo, L. S. Davis, and L. Zhang, "Weakly-supervised video summarization using variational encoder-decoder and web prior," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 184–200.

[113] H.-I. Ho, W.-C. Chiu, and Y.-C. F. Wang, "Summarizing first-person videos from third persons' points of view," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 70–85.

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

[115] L. Yang, J. Han, T. Zhao, T. Lin, D. Zhang, and J. Chen, "Background-click supervision for temporal action localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9814–9829, 2021.

[116] T. Zhao, J. Han, L. Yang, and D. Zhang, "Equivalent classification mapping for weakly supervised temporal action localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3019–3031, 2022.

[117] A. Sharghi, J. S. Laurel, and B. Gong, "Query-focused video summarization: Dataset, evaluation, and a memory network based approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4788–4797.

[118] A. Kulesza, B. Taskar *et al.*, "Determinantal point processes for machine learning," *Foundations and Trends® in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012.

[119] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," *Advances in neural information processing systems*, vol. 27, 2014.

[120] Y. Zhang, M. Kampffmeyer, X. Liang, M. Tan, and E. P. Xing, "Query-conditioned three-player adversarial network for video summarization," *arXiv preprint arXiv:1807.06677*, 2018.

[121] P. Jiang and Y. Han, "Hierarchical variational network for user-diversified & query-focused video summarization," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 2019, pp. 202–206.

[122] S. Xiao, Z. Zhao, Z. Zhang, Z. Guan, and D. Cai, "Query-biased self-attentive network for query-focused video summarization," *IEEE Transactions on Image Processing*, vol. 29, pp. 5889–5899, 2020.

[123] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, and M. Yang, "Convolutional hierarchical attention network for query-focused video summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 426–12 433.

[124] Y. Yuan, L. Ma, and W. Zhu, "Sentence specified dynamic video thumbnail generation," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 2332–2340.

[125] J.-H. Huang, L. Murn, M. Mrak, and M. Worring, "Query-based video summarization with pseudo label supervision," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 1430–1434.

[126] S. Nalla, M. Agrawal, V. Kaushal, G. Ramakrishnan, and R. Iyer, "Watch hours in minutes: Summarizing videos with user intent," in *European Conference on Computer Vision*. Springer, 2020, pp. 714–730.

[127] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[128] J.-H. Huang, C.-H. H. Yang, P.-Y. Chen, M.-H. Chen, and M. Worring, "Causalainer: Causal explainer for automatic video summarization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2629–2635.

[129] H. Li, Q. Ke, M. Gong, and 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.

[130] M. Krubiński and P. Pecina, "Mlask: multimodal summarization of video-based news articles," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 910–924.

[131] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[132] M. Narasimhan, A. Rohrbach, and T. Darrell, "Clip-it! language-guided video summarization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 988–14 000, 2021.

[133] D. Han, S. Seo, E. Park, S.-U. Nam, and N. Kwak, "Unleash the potential of clip for video highlight detection," *arXiv preprint arXiv:2404.01745*, 2024.

[134] J. Yang, P. Wei, H. Li, and Z. Ren, "Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 308–18 318.

[135] Y. Jiang, S. Luo, L. Guo, and R. Zhang, "Mct-vhd: Multi-modal contrastive transformer for video highlight detection," *Journal of Visual Communication and Image Representation*, vol. 101, p. 104162, 2024.

[136] S. Zhou, F. Zhang, R. Wang, F. Zhou, and Z. Su, "Subtask prior-driven optimized mechanism on joint video moment retrieval and highlight detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[137] Y. Xiao, Z. Luo, Y. Liu, Y. Ma, H. Bian, Y. Ji, Y. Yang, and X. Li, "Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 709–18 719.

[138] H. Hua, Y. Tang, C. Xu, and J. Luo, "V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning," *arXiv preprint arXiv:2404.12353*, 2024.

[139] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[140] G. Wu, J. Lin, and C. T. Silva, "Intentvizor: Towards generic query guided interactive video summarization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 503–10 512.

[141] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[142] Y. Zhang, Y. Liu, W. Kang, and R. Tao, "Vss-net: visual semantic self-mining network for video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[143] A. Sahu and A. S. Chowdhury, "Egocentric video co-summarization using transfer learning and refined random walk on a constrained graph," *Pattern Recognition*, vol. 134, p. 109128, 2023.

[144] M. Fei, W. Jiang, and W. Mao, "Creating personalized video summaries via semantic event detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 11, pp. 14 931–14 942, 2023.

[145] J. Wu, S.-h. Zhong, and Y. Liu, "Dynamic graph convolutional network for multi-video summarization," *Pattern Recognition*, vol. 107, p. 107382, 2020.

[146] S. Messaoud, I. Lourentzou, A. Boughoula, M. Zehni, Z. Zhao, C. Zhai, and A. G. Schwing, "Deepqamvs: Query-aware hierarchical pointer networks for multi-video summarization," in *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021, pp. 1389–1399.

[147] S. A. Ansari and A. Zafar, "Multi video summarization using query based deep optimization algorithm," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 10, pp. 3591–3606, 2023.

[148] T. T. d. S. Barbieri and R. Goularte, "Content selection criteria for news multi-video summarization based on human strategies," *International Journal on Digital Libraries*, vol. 22, no. 1, pp. 1–14, 2021.

[149] A. Sabha and A. Selwal, "Cosumnet: A video summarization-based framework for covid-19 monitoring in crowded scenes," *Artificial Intelligence in Medicine*, vol. 139, p. 102544, 2023.

[150] T. Liu, Q. Meng, A. Vlontzos, J. Tan, D. Rueckert, and B. Kainz, "Ultrasound video summarization using deep reinforcement learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 483–492.

[151] X. Zhu, J. Fan, A. K. Elmagarmid, and W. G. Aref, "Hierarchical video summarization for medical data," in *Storage and Retrieval for Media Databases 2002*, vol. 4676. SPIE, 2001, pp. 395–406.

[152] R. P. Mathews, M. R. Panicker, A. R. Hareendranathan, Y. T. Chen, J. L. Jaremko, B. Buchanan, K. V. Narayan, C. Kesavadas, and G. Mathews, "Unsupervised multi-latent space rl framework for video

summarization in ultrasound imaging," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 227–238, 2022.
[153] K. R. Raval and M. M. Goyani, "Shot segmentation and replay detection for cricket video summarization," in *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*. IEEE, 2023, pp. 933–938.
[154] P. Narwal, N. Duhan, and K. K. Bhatia, "A novel multi-modal neural network approach for dynamic and generic sports video summarization," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106964, 2023.
[155] A. Banjar, H. Dawood, A. Javed, and B. Zeb, "Sports video summarization using acoustic symmetric ternary codes and svm," *Applied Acoustics*, vol. 216, p. 109795, 2024.
[156] D. M. Davids, A. A. E. Raj, and C. S. Christopher, "Hybrid multi scale hard switch yolov4 network for cricket video summarization," *Wireless Networks*, vol. 30, no. 1, pp. 17–35, 2024.
[157] R. S. Bhat, O. Jayanth, P. Prasad, P. K. Vedurumudi, and K. Divyaprabha, "Cricket video summarization using deep learning," in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*. IEEE, 2023, pp. 1–6.
[158] 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 recognition letters*, vol. 32, no. 1, pp. 56–68, 2011.
[159] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *European conference on computer vision*. Springer, 2014, pp. 505–520.
[160] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3584–3592.
[161] Z. Ji, Y. Ma, Y. Pang, and X. Li, "Query-aware sparse coding for web multi-video summarization," *Information Sciences*, vol. 478, pp. 152–166, 2019.
[162] J.-H. Huang and M. Worring, "Query-controllable video summarization," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 242–250.
[163] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1346–1353.
[164] A. B. Vasudevan, M. Gygli, A. Volokitin, and L. Van Gool, "Query-adaptive video summarization via quality-aware relevance estimation," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 582–590.
[165] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7596–7604.
[166] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.
[167] D. Zwillinger and S. Kokoska, *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.
[168] 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 Recognition*, vol. 111, p. 107677, 2021.
[169] W. Zhu, J. Lu, Y. Han, and J. Zhou, "Learning multiscale hierarchical attention for video summarization," *Pattern Recognition*, vol. 122, p. 108312, 2022.
[170] H. Li, Q. Ke, M. Gong, and R. Zhang, "Video joint modelling based on hierarchical transformer for co-summarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3904–3917, 2022.
[171] H. Jiang and Y. Mu, "Joint video summarization and moment localization by cross-task sample transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 388–16 398.
[172] Y. Wu, Y. Wei, H. Wang, Y. Liu, S. Yang, and X. He, "Grounded image text matching with mismatched relation reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2976–2987.
[173] Y. Fan, X. Ma, R. Wu, Y. Du, J. Li, Z. Gao, and Q. Li, "Videoagent: A memory-augmented multimodal agent for video understanding," *arXiv preprint arXiv:2403.11481*, 2024.

BIOGRAPHY SECTION

Hongxi Li received the B.S. degree from Beijing Wuzi University, Beijing, China, in 2022. He will pursue the M.S. degree at the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology. His research interests include computer vision and video summarization.

Yubo Zhu received the B.S. degree in computer science from Beijing Institute of Technology, Beijing, China, in 2021. He is currently working toward the M.S. degree in computer science with the School of Computer Science and Technology, Beijing Institute of Technology. His research interests include computer vision and video summarization.

Zirui Shang received the B.S. degree in computer science in 2023 from the Beijing Institute of Technology, Beijing, China, where he is currently working toward the M.S. degree in computer technology. His research interests include vision and language, and video understanding.

Ziyi Wang received the B.S. degree from Beijing Institute of Technology, Beijing, China, in 2023. He will pursue the M.S. degree at the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology. His research interests include computer vision, domain adaptation and domain generalization.

Xinxiao Wu (Member, IEEE) is a Full Professor with the School of Computer Science, Beijing Institute of Technology. She received the B.S. degree in computer science from the Nanjing University of Information Science and Technology in 2005, and the Ph.D. degree in computer science from the Beijing Institute of Technology in 2010. From 2010 to 2011, she was a Postdoctoral Research Fellow with Nanyang Technological University, Singapore. Her research interests include machine learning, computer vision, and video analysis and understanding.