CONTRIBUTED PAPER

Video Summarization Using Deep Neural Networks: A Survey

This article provides a comprehensive survey of the existing deep-learning-based methods for generic video summarization.

By Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, *Senior Member IEEE*, and Ioannis Patras, *Senior Member IEEE*

ABSTRACT | Video summarization technologies aim to create a concise and complete synopsis by selecting the most informative parts of the video content. Several approaches have been developed over the last couple of decades, and the current state of the art is represented by methods that rely on modern deep neural network architectures. This work focuses on the recent advances in the area and provides a comprehensive survey of the existing deep-learning-based methods for generic video summarization. After presenting the motivation behind the development of technologies for video summarization, we formulate the video summarization task and discuss the main characteristics of a typical deep-learning-based analysis pipeline. Then, we suggest a taxonomy of the existing algorithms and provide a systematic review of the relevant literature that shows the evolution of the deep-learning-based video summarization technologies and leads to suggestions for future developments. We then report on protocols for the objective evaluation of video summarization algorithms, and we compare the performance of several deep-learning-based approaches. Based on the outcomes of these comparisons, as well as some documented considerations about the amount of annotated data and the suitability of evaluation protocols, we indicate potential future research directions.

KEYWORDS | Deep neural networks; evaluation protocols; summarization datasets; supervised learning; unsupervised learning; video summarization.

I. INTRODUCTION

In July 2015, YouTube revealed that it receives over 400 h of video content every single minute, which translates to 65.7 years' worth of content uploaded every day.[1] Since then, we are experiencing an even stronger engagement of consumers with both online video platforms and devices (e.g., smartphones and wearables) that carry powerful video recording sensors and allow instant uploading of the captured video on the Web. According to newer estimates, YouTube now receives 500 h of video per minute[2]; YouTube is just one of the many video hosting platforms (e.g., DailyMotion and Vimeo), social networks (e.g., Facebook, Twitter, and Instagram), and online repositories of media and news organizations that host large volumes of video content. Thus, how is it possible for someone to efficiently navigate through endless collections of videos and find the video content that she/he is looking for? The answer to this question comes not only from video retrieval technologies but also from technologies for automatic video summarization. The latter allows generating a concise synopsis that conveys the important parts of the full-length video. Given the plethora of video content on the Web,

Manuscript received October 29, 2020; revised July 2, 2021; accepted September 23, 2021. Date of current version November 1, 2021. This work was supported in part by the EU Horizon 2020 Research and Innovation Programme under Grant Agreement H2020-780656 ReTV and Grant Agreement H2020-951911 AI4Media, and in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/R026424/1. *(Corresponding author: Evlampios Apostolidis.)*

Evlampios Apostolidis is with the Information Technologies Institute/Centre for Research and Technology Hellas, GR-57001 Thessaloniki, Greece, and also with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: apostolid@iti.gr).

Eleni Adamantidou, **Alexandros I. Metsai**, and **Vasileios Mezaris** are with the Information Technologies Institute/Centre for Research and Technology Hellas, GR-57001 Thessaloniki, Greece (e-mail: adamelen@iti.gr; alexmetsai@iti.gr; bmezaris@iti.gr).

Ioannis Patras is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: i.patras@qmul.ac.uk).

Digital Object Identifier 10.1109/JPROC.2021.3117472

[1]https://www.tubefilter.com/2015/07/26/youtube-400-hours-content-every-minute/

[2]https://blog.youtube/press/

effective video summarization facilitates viewers' browsing of and navigation in large video collections, thus increasing viewers' engagement and content consumption.

The application domain of automatic video summarization is wide and includes (but is not limited to) the use of such technologies by media organizations (after integrating such techniques into their content management systems), to allow effective indexing, browsing, retrieval, and promotion of their media assets, and video sharing platforms, to improve the viewing experience, enhance viewers' engagement, and increase content consumption. In addition, video summarization that is tailored to the requirements of particular content presentation scenarios can be used for, e.g., generating trailers or teasers of movies and episodes of a TV series; presenting the highlights of an event (e.g., a sports game, a music band performance, or a public debate); and creating a video synopsis with the main activities that took place over, e.g., the last 24 h of recordings of a surveillance camera, for time-efficient progress monitoring or security purposes.

A number of surveys on video summarization have already appeared in the literature. In one of the first works, Barbieri *et al.* [1] classify the relevant bibliography according to several aspects of the summarization process, namely the targeted scenario, the type of visual content, and the characteristics of the summarization approach. In another early study, Li *et al.* [2] divide the existing summarization approaches into utility-based methods that use attention models to identify the salient objects and scenes, and structure-based methods that build on the video shots and scenes. Truong and Venkatesh [3] discuss a variety of attributes that affect the outcome of a summarization process, such as the video domain, the granularity of the employed video fragments, the utilized summarization methodology, and the targeted type of summary. Money and Agius [4] divide the bibliography into methods that rely on the analysis of the video stream, methods that process contextual video metadata, and hybrid approaches that rely on both types of the aforementioned data. Jiang *et al.* [5] discuss a few characteristic video summarization approaches that include the extraction of low-level visual features for assessing frame similarity or performing clustering-based key-frame selection; the detection of the main events of the video using motion descriptors; and the identification of the video structure using eigenfeatures. Hu *et al.* [6] classify the summarization methods into those that target minimum visual redundancy, those that rely on an object or event detection, and others that are based on multimodal integration. Ajmal *et al.* [7] similarly classify the relevant literature in clustering-based methods, approaches that rely on detecting the main events of the story, and so on. Nevertheless, all the aforementioned works (published between 2003 and 2012) report on early approaches to video summarization; they do not present how the summarization landscape has evolved over the last years and especially after the introduction of deep learning algorithms.

The more recent study of del Molino *et al.* [8] focuses on egocentric video summarization and discusses the specifications and the challenges of this task. In another recent work, Basavarajaiah and Sharma [9] provide a classification of various summarization approaches, including some recently proposed deep-learning-based methods; however, their work mainly focuses on summarization algorithms that are directly applicable to the compressed domain. Finally, the survey of Vivekraj *et al.* [10] presents the relevant bibliography based on a two-way categorization that relates to the utilized data modalities during the analysis and the incorporation of human aspects. With respect to the latter, it further splits the relevant literature into methods that create summaries by modeling the human understanding and preferences (e.g., using attention models, the semantics of the visual content, or ground-truth annotations, and machine-learning algorithms) and conventional approaches that rely on the statistical processing of low-level features of the video. Nevertheless, none of the above surveys presents, in a comprehensive manner, the current developments toward generic video summarization, which are tightly related to the growing use of advanced deep neural network architectures for learning the summarization task. As a matter of fact, the relevant research area is a very active one as several new approaches are being presented every year in highly ranked peer-reviewed journals and international conferences. In this survey, we study in detail more than 40 different deep-learning-based video summarization algorithms among the relevant works that have been proposed over the last five years. In addition, a comparison of the summarization performance reported in the most recent deep-learning-based methods against the performance reported in other more conventional approaches, e.g., [10]–[13], shows that, in most cases, the deep-learning-based methods significantly outperform more traditional approaches that rely on weighted fusion, sparse subset selection, or data clustering algorithms and represent the current state of the art in automatic video summarization. Motivated by these observations, we aim to fill this gap in the literature by presenting the relevant bibliography on deep-learning-based video summarization and also discussing other aspects that are associated with it, such as the protocols used for evaluating video summarization.

This article begins in Section II by defining the problem of automatic video summarization and presenting the most prominent types of video summary. Then, it provides a high-level description of the analysis pipeline of deep-learning-based video summarization algorithms and introduces a taxonomy of the relevant literature according to the utilized data modalities, the adopted training strategy, and the implemented learning approaches. Finally, it discusses aspects that relate to the generated summary, such as the desired properties of a static (frame-based) video summary and the length of a dynamic (fragment-based) video summary. Section III builds on the introduced

taxonomy to systematically review the relevant bibliography. A primary categorization is made according to the use or not of human-generated ground-truth data for learning, and a secondary categorization is made based on the adopted learning objective or the utilized data modalities by each different class of methods. For each one of the defined classes, we illustrate the main processing pipeline and report on the specifications of the associated summarization algorithms. After presenting the relevant bibliography, we provide some general remarks that reflect how the field has evolved, especially over the last five years, highlighting the pros and cons of each class of methods. Section IV continues with an in-depth discussion on the utilized datasets and the different evaluation protocols of the literature. Following, Section V discusses the findings of extensive performance comparisons that are based on the results reported in the relevant papers, indicates the most competitive methods in the fields of (weakly) supervised and unsupervised video summarization, and examines whether there is a performance gap between these two main types of approaches. Based on the surveyed bibliography, in Section VI, we propose potential future directions to further advance the current state of the art in video summarization. Finally, Section VII concludes this work by briefly outlining the core findings of our study.

II. PROBLEM STATEMENT

Video summarization aims to generate a short synopsis that summarizes the video content by selecting its most informative and important parts. The produced summary is usually composed of a set of representative video frames (a.k.a. video key frames) or video fragments (a.k.a. video key fragments) that have been stitched in chronological order to form a shorter video. The former type of video summary is known as video storyboard, and the latter type is known as video skim. One advantage of video skims over static sets of frames is the ability to include audio and motion elements that offer a more natural story narration and potentially enhance the expressiveness and the amount of information conveyed by the video summary. Furthermore, it is often more entertaining and interesting for the viewer to watch a skim rather than a slide show of frames [14]. On the other hand, storyboards are not restricted by timing or synchronization issues, and therefore, they offer more flexibility in terms of data organization for browsing and navigation purposes [15], [16].

A high-level representation of the typical deep-learning-based video summarization pipeline is depicted in Fig. 1. The first step of the analysis involves the representation of the visual content of the video with the help of feature vectors. Most commonly, such vectors are extracted at the frame level for all frames or for a subset of them selected via a frame-sampling strategy, e.g., processing 2 frames/s. In this way, the extracted feature vectors store information at a very detailed level and capture the dynamics of the visual content that are of high significance when selecting the video parts that form the summary. Typically, in most deep-learning-based video summarization techniques, the visual content of the video frames is represented by deep feature vectors extracted with the help of pretrained neural networks. For this, a variety of convolutional neural networks (CNNs) and deep CNNs (DCNNs) have been used in the bibliography, which includes GoogleNet (Inception V1) [17], Inception V3 [18], AlexNet [19], variations of ResNet [20], and variations of VGGnet [21]. Nevertheless, the GoogleNet appears to be the most commonly used one thus far (used in [22]–[51]). The extracted features are then utilized by a deep summarizer network, which is trained by trying to minimize an objective function or a set of objective functions.

The output of the trained deep summarizer network can be either a set of selected video frames (key frames) that form a static video storyboard or a set of selected video fragments (key fragments) that are concatenated in chronological order and form a short video skim. With respect to the generated video storyboard, this should be similar to the sets of key frames that would be selected by humans and must exhibit minimal visual redundancy. With regards to the produced video skim, this typically should be equal to or less than a predefined length L. For experimentation and comparison purposes, this is most often set as $L = p \cdot T$, where T is the video duration and p is the ratio of the summary to video length; $p = 0.15$ is a typical value, in which case the summary should not exceed 15% of the original video's duration. As a side note, the production of a video skim (which is the ultimate goal of the majority of the proposed deep-learning-based summarization algorithms) requires the segmentation of the video into consecutive and nonoverlapping fragments that exhibit visual and temporal coherence, thus offering a seamless presentation of a part of the story. Given this segmentation and the estimated frames' importance scores by the trained deep summarizer network, video-segment-level importance scores are computed by averaging the importance scores of the frames that lie within each video segment. These segment-level scores are then used to select the key fragments given the summary length L, and most methods (e.g., [22], [24]–[29], [31]–[33], [37]–[42], [44], [46], [47], [49], [50], and [52]–[54]) tackle this step by solving the Knapsack problem.

With regards to the utilized type of data, the current bibliography on deep-learning-based video summarization can be divided into the following:

1) unimodal approaches that utilize only the visual modality of the videos for feature extraction and learn summarization in a (weakly) supervised or unsupervised manner;
2) multimodal methods that exploit the available textual metadata and learn semantic-/category-driven summarization in a supervised way by increasing the relevance between the semantics of the summary and the semantics of the associated metadata or video category.

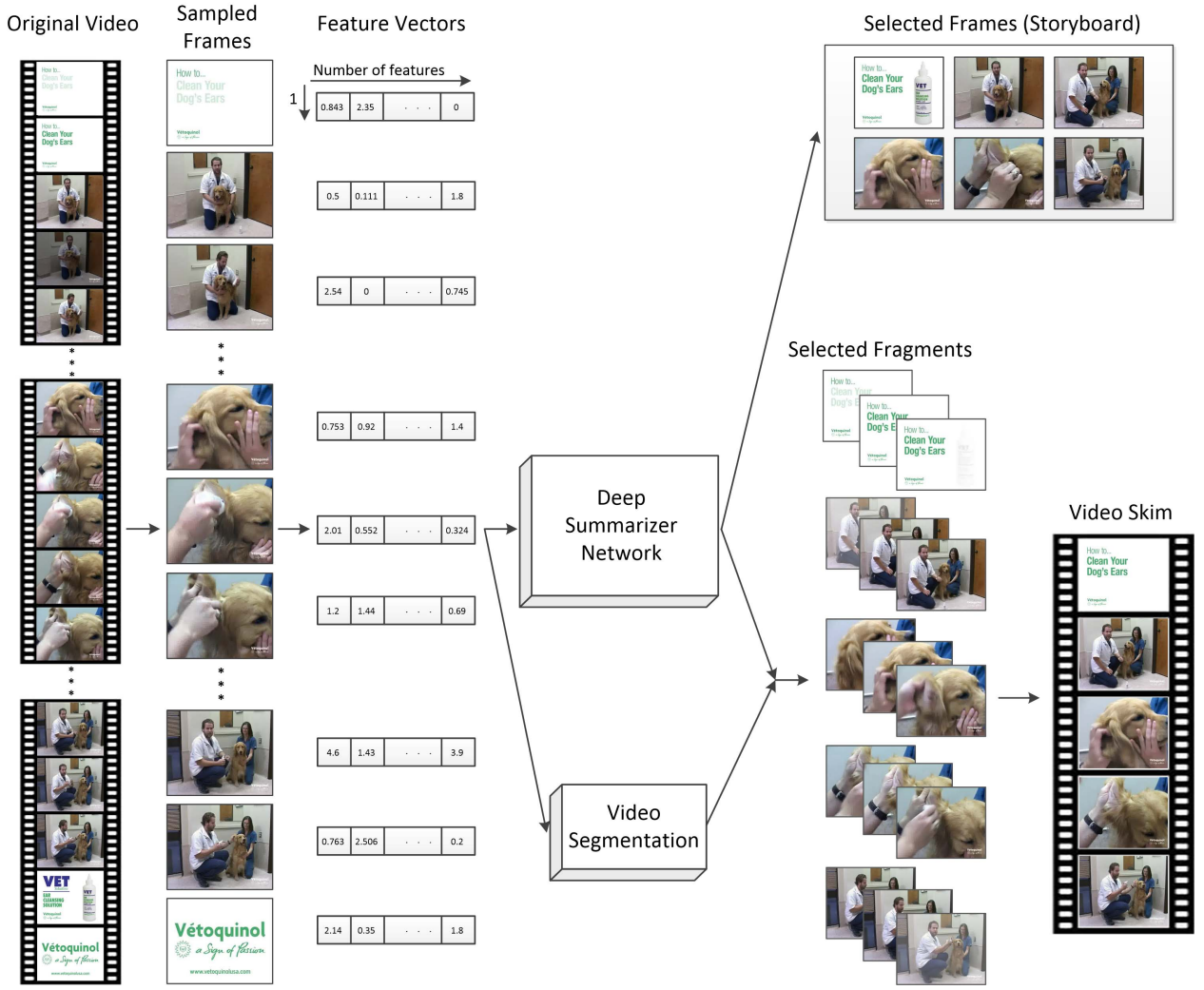


Fig. 1. ***High-level representation of the analysis pipeline of the deep-learning-based video summarization methods for generating a video storyboard and a video skim.***

Concerning the adopted training strategy, the existing deep-learning-based video summarization algorithms can be coarsely categorized in the following categories.

1) Supervised approaches that rely on datasets with human-labeled ground-truth annotations (either in the form of video summaries, as in the case of the SumMe dataset [55], or in the form of frame-level importance scores, as in the case of the TVSum dataset [56]), based on which they try to discover the underlying criterion for video frame/fragment selection and video summarization.
2) Unsupervised approaches that overcome the need for ground-truth data (whose production requires time-demanding and laborious manual annotation procedures) based on learning mechanisms that require only an adequately large collection of videos for their training.
3) Weakly supervised approaches that, similar to unsupervised approaches, aim to alleviate the need for large sets of hand-labeled data. Less-expensive weak labels are utilized with the understanding that they are imperfect compared to a full set of human annotations but can, nonetheless, be used to create strong predictive models.

Building on the above-described categorizations, a more detailed taxonomy of the relevant bibliography is depicted in Fig. 2. The penultimate layer of this arboreal illustration shows the different learning approaches that have been adopted. The leaves of each node of this layer show the utilized techniques for implementing each learning approach and contain references to the most relevant works in the bibliography. This taxonomy will be the basis for presenting the relevant bibliography in the following.

III. DEEP LEARNING APPROACHES

This section gives a brief introduction to deep learning architectures (see Section III-A) and then focuses on their application in the video summarization domain by providing a systematic review of the relevant bibliography.

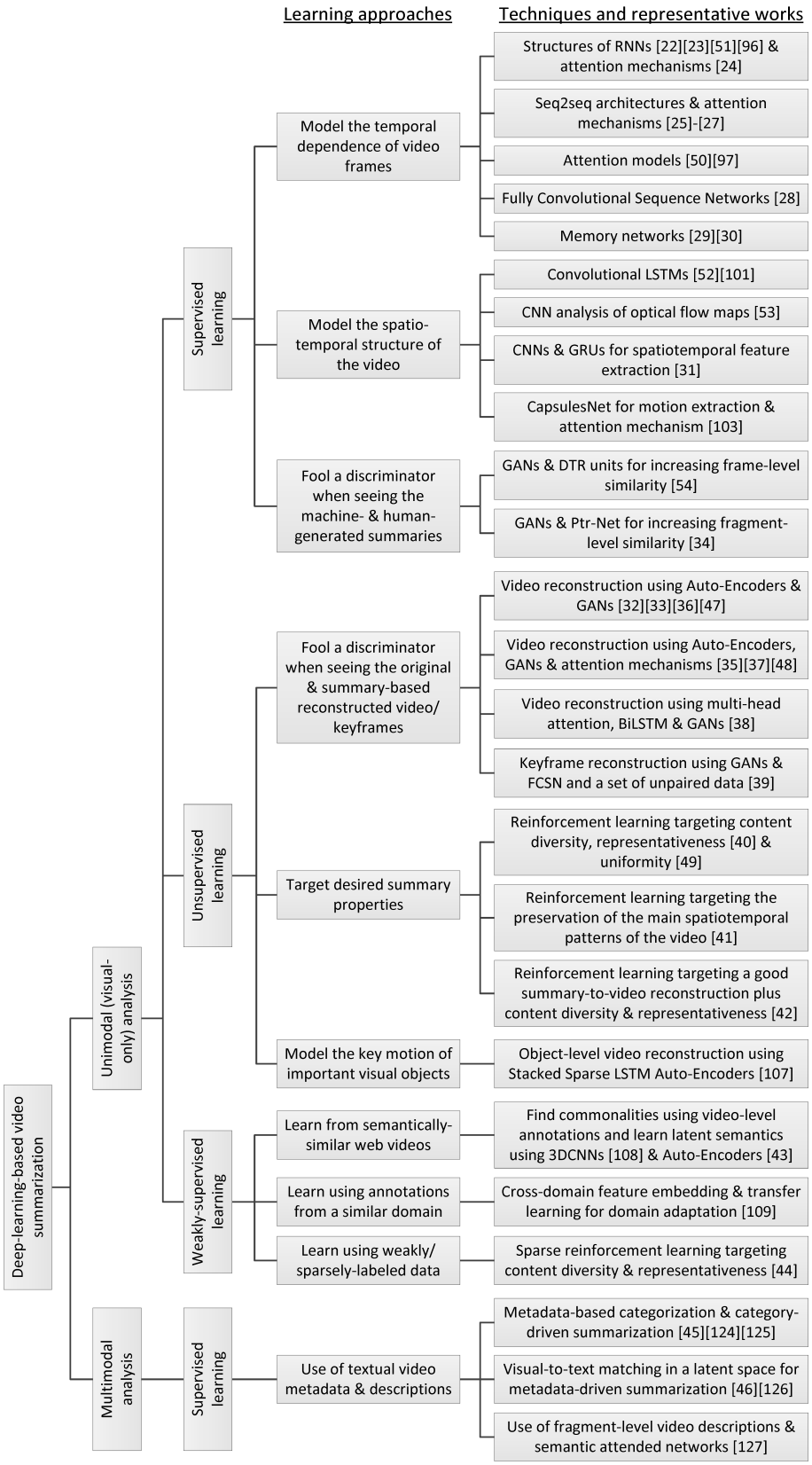


Fig. 2. ***Taxonomy of the existing deep-learning-based video summarization algorithms.***

This review starts by presenting the different classes of supervised (see Section III-B), unsupervised (see Section III-C), and weakly supervised (see Section III-D) video summarization methods, which rely solely on the analysis of the visual content. Following, it reports on multimodal approaches (see Section III-E) that process also

the available text-based metadata. Finally, it provides some general remarks (see Section III-F) that outline how the field has evolved, especially over the last five years.

A. Deep Learning Basics

Deep learning is a branch of machine learning that was fueled by the explosive growth and availability of data, and the remarkable advancements in hardware technologies. The term "deep" refers to the use of multiple layers in the network that performs nonlinear processing to learn multiple levels of data representations. Learning can be supervised, semisupervised, or unsupervised. Several deep learning architectures have been proposed thus far, which can be broadly classified in deep belief networks [57], restricted/deep Boltzmann machines [58], [59], (variational) autoencoders [60], [61], (deep) CNNs [62], recursive neural networks [63], recurrent neural networks [64], generative adversarial networks (GANs) [65], graph (convolutional) neural networks [66], and deep probabilistic neural networks [67]. For an overview of the different classes of deep learning architectures, the interested reader is referred to surveys, such as [68]–[70]. Over the last decade, deep learning architectures have been used in several applications, including natural language processing (e.g., [71] and [72]), speech recognition (e.g., [73] and [74]), medical image/video analysis (e.g., [75] and [76]), and computer vision (e.g., [20], [77]–[79]), leading to state-of-the-art results and performing in many cases comparably to a human expert. For additional information about applications of deep learning, we refer the reader to the recent surveys [80]–[86].

Nevertheless, the empirical success of deep learning architectures is associated with numerous challenges for theoreticians, which are critical to the training of deep networks. Such challenges relate to: 1) the design of architectures that are able to learn from sparse, missing, or noisy training data; 2) the use of optimization algorithms to adjust the network parameters; 3) the implementation of compact deep network architectures that can be integrated into mobile devices with restricted memory; 4) the analysis of the stability of deep networks; and 5) the explanation of the underlying mechanisms that are activated at inference time in a way that is easily understandable by humans (the relevant research domain is also known as Explainable AI). Such challenges and some suggested approaches to addressing them are highlighted in several works, e.g., [87]–[93].

B. Supervised Video Summarization

1) Learn Frame Importance by Modeling the Temporal Dependence Among Frames: Early deep-learning-based approaches cast summarization as a structured prediction problem and try to make estimates about the frames' importance by modeling their temporal dependence. As illustrated in Fig. 3, during the training phase, the summarizer gets as input the sequence of the video

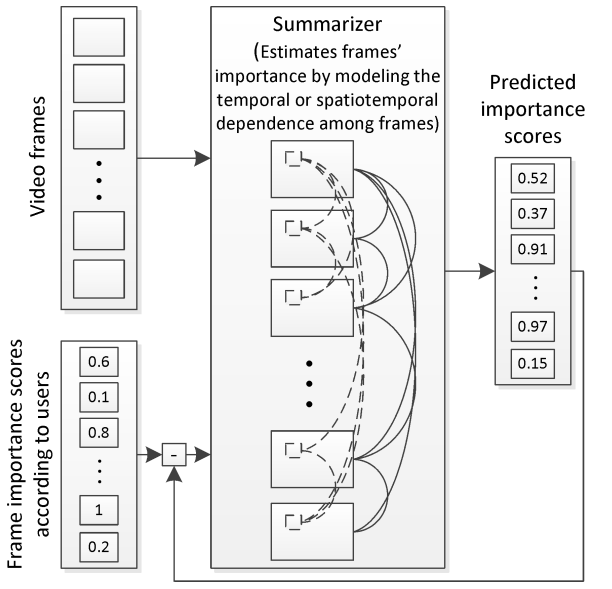


Fig. 3. ***High-level representation of the analysis pipeline of supervised algorithms that perform summarization by learning the frames' importance after modeling their temporal or spatiotemporal dependence. For the latter class of methods (i.e., modeling the spatiotemporal dependence among frames), object bounding boxes and object relations in time shown with dashed rectangles and lines are used to illustrate the extension that models both the temporal and spatial dependence among frames.***

frames and the available ground-truth data that indicate the importance of each frame according to the users' preferences. These data are then used to model the dependencies among the video frames in time (illustrated with solid arched lines) and estimate the frames' importance. The predicted importance scores are compared with the ground-truth data and the outcome of this comparison guides the training of the summarizer. The first approach to this direction, proposed by Zhang *et al.* [22], uses long short-term memory (LSTM) units [94] to model variable-range temporal dependence among video frames. Frames' importance is estimated using a multilayer perceptron (MLP), and the diversity of the visual content of the generated summary is increased based on the determinantal point process (DPP) [95]. One year later, Zhao *et al.* [23] described a two-layer LSTM architecture. The first layer extracts and encodes data about the video structure. The second layer uses this information to estimate fragment-level importance and select the key fragments of the video. In their subsequent work, Zhao *et al.* [96] integrated a component that is trained to identify the shot-level temporal structure of the video. This knowledge is then utilized for estimating importance at the shot level and producing a key-shot-based video summary. In their last work, Zhao *et al.* [51] extended the method of [23] by introducing a tensor-train embedding layer to avoid large feature-to-hidden mapping matrices. This layer is combined with a hierarchical structure of

RNNs, which operates similar to the one in [23] and captures the temporal dependence of frames that lie within manually defined video subshots (first layer) and over the different subshots of the video (second layer). The output of these layers is used for determining the probability of each subshot to be selected as a part of the video summary. Casas and Koblents [24] built on [22] and introduced an attention mechanism to model the temporal evolution of the users' interest. In the following, this information is used to estimate frames' importance and select the video key frames to build a video storyboard. In the same direction, a few methods utilized sequence-to-sequence (a.k.a. seq2seq) architectures in combination with attention mechanisms. Ji *et al.* [26] formulated video summarization as a seq2seq learning problem and proposed an LSTM-based encoder–decoder network with an intermediate attention layer. Ji *et al.* [27] introduced an extension of their summarization model from [26], which integrates a semantic preserving embedding network that evaluates the output of the decoder with respect to the preservation of the video's semantics using a tailored semantic preserving loss and replacing the previously used mean square error (MSE) loss by the Huber loss to enhance its robustness to outliers. Using the attention mechanism as the core part of the analysis and aiming to avoid the use of computationally demanding LSTMs, Fajtl *et al.* [25] presented a network for video summarization, which is composed of a soft, self-attention mechanism, and a two-layer fully connected network for regression of the frames' importance scores. Liu *et al.* [97] described a hierarchical approach that combines a generator–discriminator architecture (similar to the one in [32]) as an internal mechanism to estimate the representativeness of each shot and define a set of candidate key frames. Then, it employs a multihead attention model to further assess candidates' importance and select the key frames that form the summary. Li *et al.* [50] proposed a global diverse attention mechanism by making an adaptation of the self-attention mechanism of the transformer network [98]. This mechanism is based on a pairwise similarity matrix that contains diverse attention weights for the video frames and encodes temporal relations between every two frames in a wide range of strides. The estimated diverse attention weights are then transformed to importance scores through a regression mechanism, and these scores are compared with ground-truth annotations to learn video summarization in a supervised manner. Following another approach to model the dependence of video frames, Rochan *et al.* [28] tackled video summarization as a semantic segmentation task where the input video is seen as a 1-D image (of size equal to the number of video frames) with K channels that correspond to the K dimensions of the frames' representation vectors (either containing raw pixel values or being precomputed feature vectors). Then, they used popular semantic segmentation models, such as fully convolutional networks (FCNs) [99] and an adaptation of DeepLab [100], and built a network (called fully convolutional sequence network) for video summarization. The latter consists of a stack of convolutions with increasing effective context size as we go deeper in the network, which enables the network to effectively model long-range dependence among frames and learn frames' importance. Finally, to address issues related to the limited capacity of LSTMs, some techniques use additional memory. Feng *et al.* [29] described a deep learning architecture that stores information about the entire video in external memory and predicts each shot's importance by learning an embedding space that enables matching of each shot with the entire memory information. More recently, Wang *et al.* [30] stacked multiple LSTM and memory layers hierarchically to derive long-term temporal context and used this information to estimate the frames' importance.

2) Learn Frame Importance by Modeling the Spatiotemporal Structure of the Video: Aiming to learn how to make better estimates about the importance of video frames/fragments, some techniques pay attention to both the spatial and temporal structure of the video. Again, the summarizer gets as input the sequence of the video frames and the available ground-truth data that indicate the importance of each frame according to the users' preferences. However, extending the analysis pipeline of the previously described group of methods, it then also models the spatiotemporal dependencies among frames (shown with dashed rectangles and lines in Fig. 3). Once again, the predicted importance scores are compared with the ground-truth data, and the outcome of this comparison guides the training of the summarizer. From this perspective, Lal *et al.* [52] presented an encoder–decoder architecture with convolutional LSTMs, which models the spatiotemporal relationship among parts of the video. In addition to the estimates about the frames' importance, the algorithm enhances the visual diversity of the summary via next frame prediction and shot detection mechanisms, based on the intuition that the first frames of a shot generally have a high likelihood of being part of the summary. Yuan *et al.* [101] extracted deep and shallow features from the video content using a trainable 3-D-CNN and built a new representation through a fusion strategy. Then, they used this representation in combination with convolutional LSTMs to model the spatial and temporal structure of the video. Finally, summarization is learned with the help of a new loss function (called Sobolev loss) that aims to define a series of frame-level importance scores that are close to the series of ground-truth scores by minimizing the distance of the derivatives of these sequential data and to exploit the temporal structure of the video. Chu and Liu [53] extracted spatial and temporal information by processing the raw frames and their optical flow maps with CNNs and learned how to estimate frames' importance based on human annotations and a label distribution learning process. Elfeki and Borji [31] combined CNNs and gated recurrent units [102] (a type of RNN) to form spatiotemporal feature vectors that are

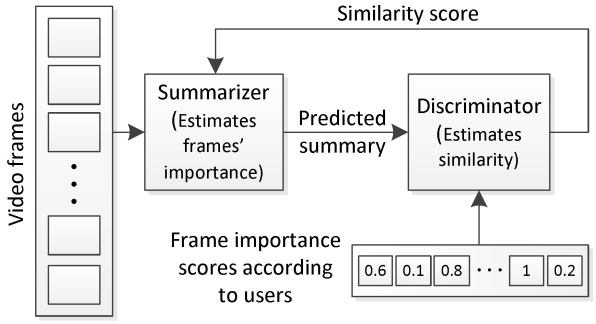


Fig. 4. ***High-level representation of the analysis pipeline of supervised algorithms that learn summarization with the help of ground-truth data and adversarial learning.***

then used to estimate the level of activity and importance of each frame. Huang and Wang [103] trained a neural network for spatiotemporal data extraction and used the extracted information to create an interframes motion curve. The latter is utilized as input to a transition effects detection method that segments the video into shots. Finally, a self-attention model exploits the human-generated ground-truth data to learn how to estimate the intrashot importance and select the key frames/fragments of the video to form a static/dynamic video summary.

3) Learn Summarization by Fooling a Discriminator When Trying to Discriminate a Machine-Generated From a Human-Generated Summary: Following a completely different approach to minimizing the distance between the machine-generated and the ground-truth summaries, a couple of methods use GANs. As presented in Fig. 4, the summarizer (which acts as the generator of the GAN) gets as input the sequence of the video frames and generates a summary by computing frame-level importance scores. The generated summary (i.e., the predicted frame-level importance scores) along with an optimal video summary for this video (i.e., frame-level importance scores according to the users' preferences) is given as input to a trainable discriminator that outputs a score that quantifies their similarity. The training of the entire summarization architecture is performed in an adversarial manner. The summarizer tries to fool the discriminator to not distinguish the predicted from the user-generated summary, and the discriminator aims to learn how to make this distinction. When the discriminator's confidence is very low (i.e., the classification error is approximately equal for both the machine- and user-generated summaries), then the summarizer is able to generate a summary that is very close to the users' expectations. In this context, Zhang *et al.* [54] proposed a method that combines LSTMs and dilated temporal relational (DTR) units to estimate temporal dependencies among frames at different temporal windows and learns summarization by trying to fool a trainable discriminator when distinguishing the machine-based summary from the ground truth and a randomly created one. In another work from the same year, Fu *et al.* [34] suggested an adversarial learning approach for (semi)supervised video summarization. The generator/summarizer is an attention-based pointer network [104] that defines the start and endpoint of each video fragment that is used to form the summary. The discriminator is a 3-D-CNN classifier that judges whether a fragment is from a ground truth or a machine-generated summary. Instead of using the typical adversarial loss, in this algorithm, the output of the discriminator is used as a reward to train the generator/summarizer based on reinforcement learning. So far, the use of GANs for supervised video summarization is limited. Nevertheless, this machine learning framework has been widely used for unsupervised video summarization, as discussed in Section III-C.

C. Unsupervised Video Summarization

1) Learn Summarization by Fooling a Discriminator When Trying to Discriminate the Original Video (or Set of Key Frames) From a Summary-Based Reconstruction of It: Given the lack of any guidance (in the form of ground-truth data) for learning video summarization, most existing unsupervised approaches rely on the rule that a representative summary ought to assist the viewer to infer the original video content. In this context, these techniques utilize GANs to learn how to create a video summary that allows a good reconstruction of the original video. The main concept of this training approach is depicted in Fig. 5. The summarizer is usually composed of a key-frame selector that estimates the frames' importance and generates a summary, and a generator that reconstructs the video based on the generated summary. It gets as input the sequence of the video frames and, through the aforementioned internal processing steps, reconstructs the original video based on the generated summary (which is represented by the predicted frame-level importance scores). The reconstructed video along with the original one is given as input to a trainable discriminator that outputs a score that quantifies their similarity. Similar

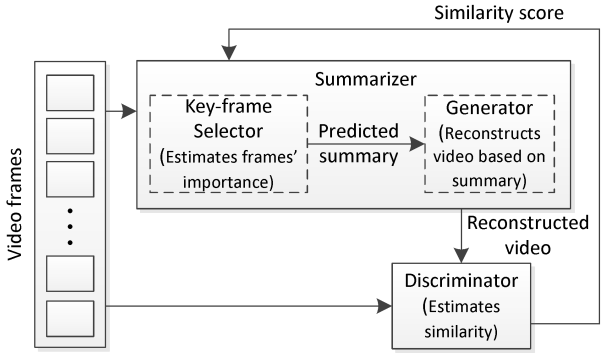


Fig. 5. ***High-level representation of the analysis pipeline of unsupervised algorithms that learn summarization by increasing the similarity between the summary and the video.***

to the supervised GAN-based methods, the training of the entire summarization architecture is performed in an adversarial manner. However, in this case, the summarizer tries to fool the discriminator when distinguishing the summary-based reconstructed video from the original one, while the discriminator aims to learn how to make the distinction. When this discrimination is not possible (i.e., the classification error is approximately equal for both the reconstructed and the original video), the summarizer is considered to be able to build a video summary that is highly representative of the overall video content. To this direction, the work of Mahasseni *et al.* [32] is the first that combines an LSTM-based key-frame selector with a variational autoencoder (VAE) and a trainable discriminator, and learns video summarization through an adversarial learning process that aims to minimize the distance between the original video and the summary-based reconstructed version of it. Apostolidis *et al.* [33] built on the network architecture of [32] and suggested a stepwise, label-based approach for training the adversarial part of the network, which leads to improved summarization performance. Yuan *et al.* [36] proposed an approach that aims to maximize the mutual information between the summary and the video using a trainable couple of discriminators and a cycle-consistent adversarial learning objective. The frame selector (bidirectional LSTM) builds a video summary by modeling the temporal dependence among frames. This summary is then forwarded to the evaluator that is composed of two GANs; the forward GAN is used to learn how to reconstruct the original video from the video summary, and the backward GAN tries to learn how to perform the backward reconstruction from the original to the summary video. The consistency between the outputs of such cycle learning is used as a measure that quantifies information preservation between the original video and the generated summary. Using this measure, the evaluator guides the frame selector to identify the most informative frames and form the video summary. In one of their subsequent works, Apostolidis *et al.* [47] embedded an actor–critic model into a GAN and formulated the selection of important video fragments (that will be used to form the summary) as a sequence generation task. The actor and the critic take part in a game that incrementally leads to the selection of the video key fragments, and their choices at each step of the game result in a set of rewards from the discriminator. The designed training workflow allows the actor and critic to discover a space of actions and automatically learn a value function (critic) and a policy for key-fragment selection (actor). In the same direction, some approaches extended the core component of the aforementioned works (i.e., the VAE-GAN architecture) by introducing tailored attention mechanisms. Jung *et al.* [35] proposed a VAE-GAN architecture that is extended by a chunk and stride network (CSNet) and a tailored difference attention mechanism for assessing the frames' dependence at different temporal granularities when selecting the video key frames. In their next work, Jung *et al.* [48] introduced another approach for estimating frames' importance, which uses a self-attention mechanism (similar to the one integrated into the transformer network [98]) in combination with an algorithm for modeling the relative position between frames. The frame sequence is decomposed into equally sized, nonoverlapping groups of consecutive, and neighboring frames (selected using a constant sampling step) to capture both the local and global interdependencies between video frames. The proposed approach was considered as a strategy for estimating frames' importance, and its effectiveness was evaluated after being integrated into the network architecture of [35]. Apostolidis *et al.* [37] introduced a variation of their previous work [33] that replaces the VAE with a deterministic attention autoencoder for learning an attention-driven reconstruction of the original video, which subsequently improves the key-fragment selection process. He *et al.* [38] presented a self-attention-based conditional GAN. The generator produces weighted frame features and predicts frame-level importance scores, while the discriminator tries to distinguish between the weighted and the raw frame features. A conditional feature selector is used to guide the GAN model to focus on more important temporal regions of the whole video frames, while long-range temporal dependencies along the whole video sequence are modeled by a multihead self-attention mechanism. Finally, building on a generator–discriminator mechanism, Rochan and Wang [39] proposed an approach that learns video summarization from unpaired data based on an adversarial process that relies on GANs and a fully convolutional sequence network (FCSN) encoder–decoder. The model of [39] aims to learn a mapping function of a raw video to a human-like summary such that the distribution of the generated summary is similar to the distribution of human-created summaries, while content diversity is forced by applying a relevant constraint on the learned mapping function.

2) Learn Summarization by Targeting Specific Desired Properties for the Summary: Aiming to deal with the unstable training [40] and the restricted evaluation criteria of GAN-based methods (that mainly focus on the summary's ability to lead to a good reconstruction of the original video), some unsupervised approaches perform summarization by targeting specific properties of an optimal video summary. To this direction, they utilize the principles of reinforcement learning in combination with handcrafted reward functions that quantify the existence of desired characteristics in the generated summary. As presented in Fig. 6, the summarizer gets as input the sequence of the video frames and creates a summary by predicting frame-level importance scores. The created (predicted) summary is then forwarded to an evaluator, which is responsible to quantify the existence of specific desired characteristics with the help of handcrafted reward functions. The computed score(s) are then combined to form an overall reward value, which is finally used to guide the

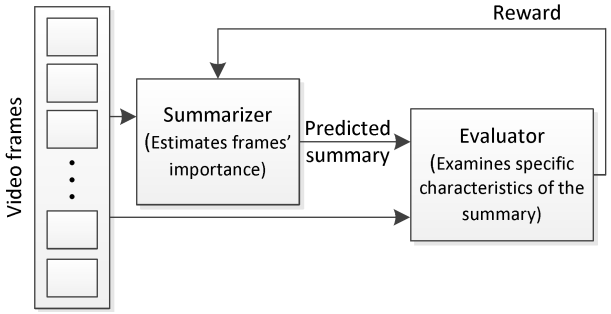


Fig. 6. ***High-level representation of the analysis pipeline of supervised algorithms that learn summarization based on handcrafted rewards and reinforcement learning.***

training of the summarizer. The first work in this direction, proposed by Zhou and Qiao [40], formulates video summarization as a sequential decision-making process and trains a summarizer to produce diverse and representative video summaries using a diversity-representativeness reward. The diversity reward measures the dissimilarity among the selected key frames, and the representativeness reward computes the distance (expressing the visual resemblance) of the selected key frames from the remaining frames of the video. Building on this method, Yaliniz and Ikizler-Cinbis [49] presented another reinforcement-learning-based approach that considers also the uniformity of the generated summary. The temporal dependence among frames is modeled using independently recurrent neural networks (IndRNNs) [105] activated by a leaky rectified linear unit (ReLU) function; in this way, Yaliniz and Ikizler-Cinbis [49] try to overcome identified issues of LSTMs with regards to decaying, vanishing and exploding gradients, and better learn long-term dependencies over the sequence of video frames. Moreover, besides using rewards associated with the representativeness and diversity of the video summary, to avoid redundant jumps between the selected video fragments that form the summary, Yaliniz and Ikizler-Cinbis [49] added a uniformity reward that aims to enhance the coherence of the generated summary. Gonuguntla *et al.* [41] built a method that utilizes temporal segment networks (proposed in [106] for action recognition in videos) to extract spatial and temporal information about the video frames and trains the summarizer through a reward function that assesses the preservation of the video's main spatiotemporal patterns in the produced summary. Finally, Zhao *et al.* [42] presented a mechanism for both video summarization and reconstruction. Video reconstruction aims to estimate the extent to which the summary allows the viewer to infer the original video (similar to some of the above-presented GAN-based methods), and video summarization is learned based on the reconstructor's feedback and the output of trained models that assess the representativeness and diversity of the visual content of the generated summary.

3) Build Object-Oriented Summaries by Modeling the Key Motion of Important Visual Objects: Building on a different basis, Zhang *et al.* [107] developed a method that focuses on the preservation in the summary of the underlying fine-grained semantic and motion information of the video. For this, it performs a preprocessing step that aims to find important objects and their key motions. Based on this step, it represents the whole video by creating supersegmented object motion clips. Each one of these clips is then given to the summarizer, which uses an online motion autoencoder model (stacked sparse LSTM autoencoder) to memorize past states of object motions by continuously updating a tailored recurrent autoencoder network. The latter is responsible for reconstructing object-level motion clips, and the reconstruction loss between the input and the output of this component is used to guide the training of the summarizer. Based on this training process, the summarizer is able to generate summaries that show the representative objects in the video and the key motions of each of these objects.

D. Weakly Supervised Video Summarization

Weakly supervised video summarization methods try to mitigate the need for extensive human-generated ground-truth data, similar to unsupervised learning methods. Instead of not using any ground-truth data, they use less-expensive weak labels (such as video-level metadata for video categorization and category-driven summarization, or ground-truth annotations for a small subset of video frames for learning summarization through sparse reinforcement learning and tailored reward functions) under the main hypothesis that these labels are imperfect compared to a full set of human annotations but can, nonetheless, allow the training of effective summarization models. We avoid the illustration of a typical analysis pipeline for this class of methods, as there is limited overlap in the way that these methods conceptualize the learning of the summarization task. The first approach that adopts an intermediate way between fully supervised and fully unsupervised learning for video summarization was described by Panda *et al.* [108]. This approach uses video-level metadata (e.g., the video title "A man is cooking") to define a categorization of videos. Then, it leverages multiple videos of a category and extracts 3-D-CNN features to automatically learn a parametric model for categorizing new (unseen during training) videos. Finally, it adopts the learned model to select the video segments that maximize the relevance between the summary and the video category. Panda *et al.* [108] investigated different ways to tackle issues related to the limited size of available datasets that include cross-dataset training, the use of web-crawled videos, and data augmentation methods for obtaining sufficient training data. Building on the concept of learning summarization from semantically similar videos, Cai *et al.* [43] suggested a weakly supervised setting of learning summarization models from

a large number of web videos. Their architecture combines a VAE that learns the latent semantics from web videos and a sequence encoder–decoder with an attention mechanism that performs summarization. The decoding part of the VAE aims to reconstruct the input videos using samples from the learned latent semantics, while the most important video fragments are identified through the soft attention mechanism of the encoder–decoder network, where the attention vectors of raw videos are obtained by integrating the learned latent semantics from the collected web videos. The overall architecture is trained by a weakly supervised semantic matching loss to learn the topic-associated summaries. Ho *et al.* [109] proposed a deep learning framework for summarizing first-person videos; however, we report on this method here, as it is also evaluated on a dataset used to assess generic video summarization methods. Given the observation in [109] that the collection of a sufficiently large amount of fully annotated first-person video data with ground-truth annotations is a difficult task, Ho *et al.* built an algorithm that exploits the principles of transfer learning and uses annotated third-person videos (which, as argued in [109], can be found more easily) to learn how to summarize first-person videos. The algorithm performs cross-domain feature embedding and transfer learning for domain adaptation (across third- and first-person videos) in a semi-supervised manner. In particular, training is performed based on a set of third-person videos with fully annotated highlight scores and a set of first-person videos where only a small portion of them comes with ground-truth scores. Finally, Chen *et al.* [44] utilized the principles of reinforcement learning to build and train a summarization method based on a limited set of human annotations and a set of handcrafted rewards. The latter relates to the similarity between the machine- and human-selected fragments, as well as to specific characteristics of the created summary (e.g., its representativeness). More specifically, this method applies a hierarchical key-fragment selection process that is divided into subtasks. Each task is learned through sparse reinforcement learning (thus avoiding the need for exhaustive annotations about the entire set of frames and using annotations only for a subset of frames), and the final summary is formed based on rewards about its diversity and representativeness.

E. Multimodal Approaches

A number of works investigated the potential of exploiting additional modalities (besides the visual stream) for learning summarization, such as the audio stream, the video captions or ASR transcripts, any available textual metadata (video title and/or abstract), or other contextual data (e.g., viewers' comments). Several of these multimodal approaches were proposed before the so-called "deep learning era," targeting either generic or domain-/task-specific video summarization. Some indicative and recent examples can be found in [110]–[117].

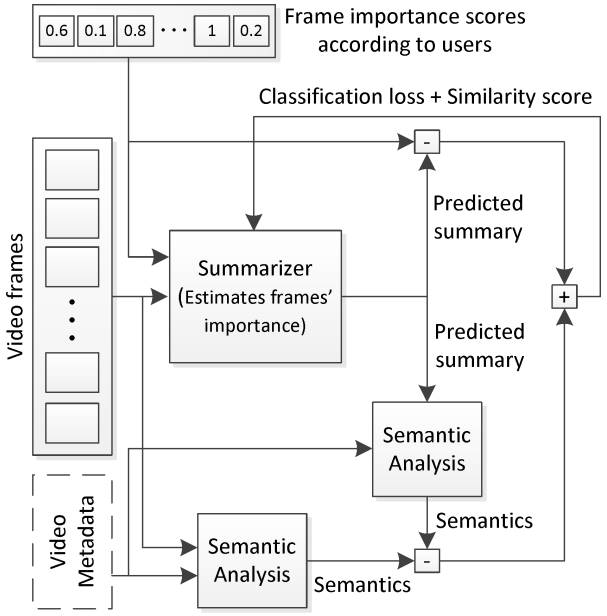


Fig. 7. ***High-level representation of the analysis pipeline of supervised algorithms that perform semantic-/category-driven summarization.***

Addressing the video summarization task from a slightly different perspective, other multimodal algorithms generate a text-based summary of the video [118]–[120] and extend this output by extracting one representative key frame [121]. A multimodal deep-learning-based algorithm for summarizing videos of soccer games was presented in [122], while another multimodal approach for key-frame extraction from first-person videos that exploits sensor data was described in [123]. Nevertheless, all these works are outside the scope of this survey, which focuses on deep-learning-based methods for generic video summarization, i.e., methods that learn summarization with the help of deep neural network architectures and/or the visual content are represented by deep features.

The majority of multimodal approaches that are within the focus of this study tackles video summarization by utilizing also textual video metadata, such as the video title and description. As depicted in Fig. 7, during training, the summarizer gets as input: 1) the sequence of the video frames that need to be summarized; 2) the relevant ground-truth data (i.e., frame-level importance scores according to the users); and 3) the associated video metadata. In the following, it estimates the frames' importance and generates (predicts) a summary. Then, the generated summary is compared with the ground-truth summary and the video metadata. The former comparison produces a similarity score. The latter comparison involves the semantic analysis of the summary and the video metadata. Given the output of this analysis, most algorithms try to minimize the distance of the generated representations in a learned latent space or the classification loss when the aim is to define a summary that

maintains the core characteristics of the video category. The combined outcome of this assessment is finally used to guide the training of the summarizer. In this context, Zhou *et al.* [45] and Song *et al.* [124] proposed methods that learn category-driven summarization by rewarding the preservation of the core parts found in video summaries from the same category (e.g., the main parts of a wedding ceremony when summarizing a wedding video). In the same direction, Lei *et al.* [125] presented a method that uses action classifiers that have been trained with video-level labels to perform action-driven video fragmentation and labeling; then, this method extracts a fixed number of key frames and applies reinforcement learning to select the ones with the highest categorization accuracy, thus performing category-driven summarization. Building on the idea of exploiting contextual data, Yuan *et al.* [46] and Otani *et al.* [126] suggested methods that define a video summary by maximizing its relevance with the available video metadata, after projecting both the visual and the textual information in a common latent space. Finally, approaching the summarization task from a different perspective, Wei *et al.* [127] introduced an approach that initially learns a visual-to-text mapping using fragment-level human descriptions of the training videos. In the following, for a given training sample, it applies the learned mapping to describe the content of both the original video and the generated video summary and learns summarization by minimizing the distance of these textual descriptions and the distance between the machine and the human summaries, using semantic attended networks. However, most of these methods examine only the visual cues without considering the sequential structure of the video and the temporal dependence among frames.

F. General Remarks on Deep Learning Approaches

Based on the review of the relevant bibliography, we saw that early deep-learning-based approaches for video summarization utilize combinations of CNNs and RNNs. The former is used as pretrained components (e.g., CNNs trained on ImageNet for visual concept detection) to represent the visual content, and the latter (mostly LSTMs) is used to model the temporal dependence among video frames. The majority of these approaches are supervised and try to learn how to make estimates about the importance of video frames/fragments based on human-generated ground-truth annotations. Architectures of RNNs can be used in the typical or in a hierarchical form [22], [23], [51], [96] to also model the temporal structure and utilize this knowledge when selecting the video fragments of the summary. In some cases, such architectures are combined with attention mechanisms to model the evolution of the users' interest [24], [26], [27] or extended by memory networks to increase the memorization capacity of the architecture and capture long-range temporal dependencies among parts of the video [29], [30]. Alternatively, some works [25], [50] avoid the use of computationally demanding RNNs, and instead, they model the frames' dependencies with the help of learnable similarity-based attention mechanisms. Going one step further, a group of techniques tries to learn importance by paying attention to both the spatial and temporal structure of the video, using convolutional LSTMs [52], [101], optical flow maps [53], combinations of CNNs and RNNs [31], or motion extraction mechanisms [103]. Following a different approach, a couple of supervised methods learn how to generate video summaries that are aligned with the human preferences with the help of GANs [34], [54]. Finally, multimodal algorithms extract the high-level semantics of the visual content using pretrained CNNs/DCNNs and learn summarization in a supervised manner by maximizing the semantic similarity among the visual summary and the contextual video metadata [46], [126], the video category [45], [125], or human descriptions of the video content [127]. However, the latter methods focus mostly on the visual content and disregard the sequential nature of the video, which is essential when summarizing the presented story.

To train a video summarization network in a fully unsupervised manner, the use of GANs seems to be the central direction thus far [32], [33], [35]–[39], [47]. The main intuition behind the use of GANs is that the produced summary should allow the viewer to infer the original video, and thus, the unsupervised GAN-based methods are trying to build a summary that enables a good reconstruction of the original video. In most cases, the generative part is composed of an LSTM that estimates the frames' importance according to their temporal dependence, thus indicating the most appropriate video parts for the summary. Then, the reconstruction of the video based on the predicted summary is performed with the help of autoencoders [32], [33], [36] that, in some cases, are combined with tailored attention mechanisms [35], [37]. Alternatively, the selection of the most important frames or fragments can be assisted by the use of actor–critic models [47] or transformer-like self-attention mechanisms [48]. Another, but the less popular, approach for unsupervised video summarization is the use of reinforcement learning in combination with handcrafted rewards about specific properties of the generated summary. These rewards usually aim to increase the representativeness, diversity [40], and uniformity [49] of the summary, retain the spatiotemporal patterns of the video [41], or secure a good summary-based reconstruction of the video [42]. Finally, one unsupervised method learns how to build object-oriented summaries by modeling the key motion of important visual objects using a stacked sparse LSTM autoencoder [107].

Last but not least, a few weakly supervised methods have also been proposed. These methods learn video summarization by exploiting the semantics of the video metadata [108] or the summary structure of semantically similar web videos [43], exploiting annotations from a similar domain and transferring the gained knowledge

via cross-domain feature embedding and transfer learning techniques [109], or using weakly/sparsely labeled data under a reinforcement learning framework [44].

With respect to the potential of deep-learning-based video summarization algorithms, we argue that, despite the fact that, currently, the major research direction is toward the development of supervised algorithms, the exploration of the learning capability of fully unsupervised and semisupervised/weakly supervised methods is highly recommended. The reasoning behind this claim is based on the fact that: 1) the generation of ground-truth training data (summaries) can be an expensive and laborious process; 2) video summarization is a subjective task, and thus, multiple different summaries can be proposed for a video from different human annotators; and 3) these ground-truth summaries can vary a lot, thus making it hard to train a method with the typical supervised training approaches. On the other hand, unsupervised video summarization algorithms overcome the need for ground-truth data as they can be trained using only an adequately large collection of original, full-length videos. Moreover, unsupervised and semisupervised/weakly supervised learning allows to easily train different models of the same deep network architecture using different types of video content (TV shows and news) and user-specified rules about the content of the summary, thus facilitating the domain adaptation of video summarization. Given the above, we believe that unsupervised and semisupervised/weakly supervised video summarizations have great advantages, and thus, their potential should be further investigated.

IV. EVALUATING VIDEO SUMMARIZATION

A. Datasets

Four datasets prevail in the video summarization bibliography: SumMe [55], TVSum [56], OVP [128], and YouTube [128]. SumMe consists of 25 videos of 1–6-min duration, with diverse video contents, captured from both first- and third-person views. Each video has been annotated by 15–18 users in the form of key fragments and, thus, is associated with multiple fragment-level user summaries that have a length between 5% and 15% of the initial video duration. TVSum consists of 50 videos of 1–11-min duration, containing video content from ten categories of the TRECVid MED dataset. The TVSum videos have been annotated by 20 users in the form of shot- and frame-level importance scores (ranging from 1 to 5). OVP and YouTube both contain 50 videos, whose annotations are sets of key frames, produced by five users. The video duration ranges from 1 to 4 min for OVP and from 1 to 10 min for YouTube. Both datasets are comprised of videos with diverse video content, such as documentaries, educational, ephemeral, historical, and lecture videos (OVP dataset) and cartoons, news, sports, commercials, TV shows, and home videos (YouTube dataset). Given the size of each of these datasets, we argue that there is a lack of large-scale annotated datasets that could be useful for improving the training of complex supervised deep learning architectures.

Some less-commonly used datasets for video summarization are CoSum [129], MED-summaries [130], Video Titles in the Wild (VTW) [131], League of Legends (LoL) [132], and FVPSum [109]. CoSum has been created to evaluate video cosummarization. It consists of 51 videos that were downloaded from YouTube using ten query terms that relate to the video content of the SumMe dataset. Each video is approximately 4-min long, and it is annotated by three different users who have selected sets of key fragments. The MED-summaries dataset contains 160 annotated videos from the TRECVID 2011 MED dataset; 60 videos form the validation set (from 15 event categories), and the remaining 100 videos form the test set (from ten event categories), with most of them being 1–5 min long. The annotations come as one set of importance scores, averaged over one to four annotators. As far as the VTW dataset is concerned, it includes 18 100 open domain videos, with 2000 of them being annotated in the form of subshot level highlight scores. The videos are user-generated, untrimmed videos that contain a highlight event and have an average duration of 1.5 min. Regarding LoL, it has 218 long videos (30–50 min), displaying game matches from the North American League of Legends Championship Series (NALCS). The annotations derive from a YouTube channel that provides community-generated highlights (videos with a duration of 5–7 min). Therefore, one set of key fragments is available for each video. Finally, FPVSum is a first-person video summarization dataset. It contains 98 videos (more than 7-h total duration) from 14 categories of GoPro viewer-friendly videos. For each category, about 35% of the video sequences are annotated with ground-truth scores by at least 10 users, while the remaining are viewed as unlabeled examples. The main characteristics of each of the above-discussed datasets are briefly presented in Table 1.

It is worth mentioning that, in this work, we focus only on datasets that are fit for evaluating video summarization methods, namely, datasets that contain ground-truth annotations regarding the summary or the frame-/fragment-level importance of each video. Other datasets might be also used by some works for network pretraining purposes, but they do not concern this work. Table 2 summarizes the datasets utilized by the deep-learning-based approaches for video summarization. From this table, it is clear that the SumMe and TVSum datasets are, by far, the most commonly used ones. OVP and YouTube are also utilized in a few works but mainly for data augmentation purposes.

B. Evaluation Protocols and Measures

Several approaches have been proposed in the literature for evaluating the performance of video summarization. A categorization of these approaches, along with a brief presentation of their main characteristics, is provided in Table 3. In the sequel, we discuss in more detail these

Table 1 Datasets for Video Summarization and Their Main Characteristics

Dataset	# of videos	duration (min)	content	type of annotations	# of annotators per video
SumMe [55]	25	1 - 6	holidays, events, sports	multiple sets of key-fragments	15 - 18
TVSum [56]	50	1 - 11	news, how-to's, user-generated, documentaries (10 categories - 5 videos each)	multiple fragment-level scores	20
OVP [128]	50	1 - 4	documentary, educational, ephemeral, historical, lecture	multiple sets of key-frames	5
Youtube [128]	50	1 - 10	cartoons, sports, tv-shows, commercial, home videos	multiple sets of key-frames	5
CoSum [129]	51	~ 4	holidays, events, sports (10 categories)	multiple sets of key-fragments	3
MED [130]	160	1 - 5	15 categories of various genres	one set of imp. scores	1 - 4
VTW [131]	2000	1.5 (avg)	user-generated videos that contain a highlight event	sub-shot level highlight scores	-
LoL [132]	218	30 - 50	matches from a League of Legends tournament	one set of key-fragments	1
FPVSum [109]	98	4.3 (avg)	first-person videos (14 categories)	multiple frame-level scores	10

evaluation protocols in chronological order to show the evolution of ideas on the assessment of video summarization methods.

1) Evaluating Video Storyboards: Early video summarization techniques created a static summary of the video content with the help of representative key frames. First

Table 2 Datasets Used by Each Deep-Learning-Based Method for Evaluating Video Summarization Performance

Method	SumMe	TVSum	OVP	Youtube	CoSum	MED	VTW	LoL	FPVSum
vsLSTM (2016) [22]	✓	✓	✓	✓					
dppLSTM (2016) [22]	✓	✓	✓	✓					
H-RNN (2017) [23]	✓	✓				✓	✓		
SUM-GAN (2017) [32]	✓	✓	✓	✓					
DeSumNet (2017) [108]		✓			✓				
VS-DSF (2017) [126]	✓								
HSA-RNN (2018) [96]	✓	✓			✓		✓		
SUM-FCN (2018) [28]	✓	✓	✓	✓					
MAVS (2018) [29]	✓	✓							
DR-DSN (2018) [40]	✓	✓	✓	✓					
Online Motion-AE (2018) [107]	✓	✓		✓					
FPVSF (2018) [109]	✓	✓							✓
VESD (2018) [43]		✓			✓				
DQSN (2018) [45]		✓			✓				
SASUM (2018) [127]	✓	✓		✓					
vsLSTM+Att (2019) [24]	✓	✓	✓	✓					
dppLSTM+Att (2019) [24]	✓	✓	✓	✓					
VASNet (2019) [25]	✓	✓	✓	✓					
H-MAN (2019) [97]	✓	✓	✓	✓					
SMN (2019) [30]	✓	✓	✓	✓					
CRSum (2019) [101]	✓	✓					✓		
SMLD (2019) [53]	✓	✓							
ActionRanking (2019) [31]	✓	✓	✓	✓					
DTR-GAN (2019) [54]		✓							
Ptr-Net (2019) [34]	✓	✓		✓				✓	
SUM-GAN-sl (2019) [33]	✓	✓							
CSNet (2019) [35]	✓	✓	✓	✓					
Cycle-SUM (2019) [36]	✓	✓							
ACGAN (2019) [38]	✓	✓	✓	✓					
UnpairedVSN (2019) [39]	✓	✓	✓	✓					
EDSN (2019) [41]	✓	✓							
WS-HRL (2019) [44]	✓	✓	✓	✓					
DSSE (2019) [46]		✓							
A-AVS (2020) [26]	✓	✓	✓	✓					
M-AVS (2020) [26]	✓	✓	✓	✓					
DASP (2020) [27]	✓	✓		✓					
SF-CVS (2020) [103]	✓	✓		✓					
SUM-GAN-AAE (2020) [37]	✓	✓							
PCDL (2020) [42]	✓	✓	✓	✓					
TTH-RNN (2020) [51]	✓	✓				✓	✓		
GL-RPE (2020) [48]	✓	✓							
AC-SUM-GAN (2021) [47]	✓	✓							
SUM-Ind_{LU} (2021) [49]	✓	✓							
SUM-GDA (2021) [50]	✓	✓	✓	✓			✓		

Table 3 Proposed Protocols for the Evaluation of Video Summarization Methods

Qualitative evaluation of video storyboards based on user studies	
Relevant works	**Adopted criteria**
Avila et al. (2008) [133]	Relevance of each key-frame with the video content and redundant or missing information in the key-frame set
Ejaz et al. (2014) [134]	Informativeness and enjoyability of the key-frame set
Quantitative evaluation of video storyboards based on ground-truth annotations	
Relevant works	**Used measures**
Chasanis et al. (2008) [135]	Fidelity (min. distance of a frame from the key-frame set) and Shot Reconstruction Degree (level of reconstruction of the entire frame sequence using the key-frame set)
Avila et al. (2011) [128] (also used in [136]–[139])	Overlap between machine- and multiple user-generated key-frame-based summaries measured by Accuracy and Error Rates (known as "Comparison of User Summaries")
Mahmoud et al. (2013) [140] (also used in [141]–[144])	Overlap between machine- and multiple user-generated key-frame-based summaries measured by Precision, Recall, F-Score
Quantitative evaluation of video skims based on ground-truth annotations	
Relevant works	**Used measures**
Gygli et al. (2014) [55] (also used in [11], [22], [23], [25], [26], [30], [31] [33], [35], [37]–[40], [42], [45], [47], [48], [50]–[52] [53], [56], [96], [97], [126], [127], [145], [146])	Overlap between machine- and multiple user-generated key-fragment-based summaries measured by Precision, Recall, F-Score
Mahasseni et al. (2017) [32] (also used in [34], [36], [49], [54])	Overlap between machine- and single ground-truth key-fragment-based summary measured by Precision, Recall, F-Score
Otani et al. (2019) [147] (also used in [44], [48])	Alignment between machine- and multiple user-generated series of frame-level importance scores using the Kendall and Spearman rank correlation coefficients
Apostolidis et al. (2020) [148]	Extension of the one from Gygli et al. [55]. The performance of the machine-based summarizer is divided by the performance of a random summarizer (known as "Performance Over Random")

attempts toward the evaluation of the created key-frame-based summaries were based on user studies that made use of human judges for evaluating the resulting quality of the summaries. Judgment was based on specific criteria, such as the relevance of each key frame with the video content, redundant or missing information [133], or the informativeness and enjoyability of the summary [134]. Although this can be a useful way to evaluate results, the procedure of evaluating each resulting summary by users can be time-consuming, and the evaluation results cannot be easily reproduced or used in future comparisons. To overcome the deficiencies of user studies, other works evaluate their key-frame-based summaries using objective measures and ground-truth summaries. In this context, Chasanis *et al.* [135] estimated the quality of the produced summaries using the fidelity measure [149] and the shot reconstruction degree criterion [150]. A different approach, termed "Comparison of User Summaries," was proposed in [128] and evaluates the generated summary according to its overlap with predefined key-frame-based user summaries. Comparison is performed at a key-frame-basis, and the quality of the generated summary is quantified by computing the accuracy and error rates based on the number of matched and nonmatched key frames, respectively. This approach was also used in [136]–[139]. A similar methodology was followed in [140]. Instead of accuracy and error rates, in [140], the evaluation relies on the well-known precision, recall, and F-Score measures. This protocol was also used in the supervised video summarization approach of [141], while a variation of it was used in [142]–[144]. In the latter case, in addition to their visual similarity, two frames are considered a match only if they are temporally no more than a specific number of frames apart.

2) Evaluating Video Skims: Most recent algorithms tackle video summarization by creating a dynamic video summary (video skim). For this, they select the most representative video fragments and join them in a sequence to form a shorter video. The evaluation methodologies of these works assess the quality of video skims according to their alignment with human preferences. Contrary to the early approaches for video storyboard generation that utilized qualitative evaluation methodologies, these works extend the evaluation protocols of the late video storyboard generation approaches and perform the evaluation using ground-truth data and objective measures. A first attempt was made in [55], where an evaluation approach along with the SumMe dataset for video summarization was introduced. According to this approach, the videos are first segmented into consecutive and nonoverlapping fragments in order to enable matching between key-fragment-based summaries (i.e., to compare the user-generated with the automatically created summary). Then, based on the scores computed by a video summarization algorithm for the fragments of a given video, an optimal subset of them (key fragments) is selected and forms the summary. The alignment of this summary with the user summaries for this video is evaluated by computing F-Score in a pairwise manner. In particular, the F-Score for the summary of the ith video is computed as follows:

$$F_i = \frac{1}{N_i} \sum_{j=1}^{N_i} 2 \frac{P_{i,j} \, R_{i,j}}{P_{i,j} + R_{i,j}} \tag{1}$$

where N_i is the number of available user-generated summaries for the ith test video, $P_{i,j}$ and $R_{i,j}$ are the precision and recall against the jth user summary, and they are both computed on a per-frame basis. This methodology was adopted also in [11], [56], and [126]. The latter work introduced another dataset, called TVSum. As in [55], the shots of the videos of the TVSum dataset were defined through automatic video segmentation. Based on the

results of a video summarization algorithm, the computed (frame or fragment level) scores are used to define the sequence of selected video fragments and produce the summary. Similar to [55], for a given video of the TVSum dataset, the agreement of the created summary with the user summaries is quantified by F-Score.

The evaluation approach and benchmark datasets of [55] and [56] were jointly used to evaluate the summarization performance in [22]. After defining a new segmentation for the videos of both datasets, Zhang *et al.* [22] evaluated the efficiency of their method on both datasets based on the multiple user-generated summaries for each video. Moreover, they documented the needed conversions from frame-level importance scores to key-fragment-based summaries in the Supplementary Material of [22]. The typical settings about the data split into training and testing (80% for training and 20% for testing) and the target summary length ($\leq$15% of the video duration) were used, and the evaluation was based on F-Score. Experiments were conducted five times, and the authors report the average performance and the standard deviation (STD). The above-described evaluation protocol—with slight variations that relate to the number of iterations using different randomly created splits of the data (five-splits; ten-splits; "few"-splits; and fivefold cross validation), the way that the computed F-Scores from the pairwise comparisons with the different user summaries are taken under consideration (maximum value is kept for SumMe according to [11]; average value is kept for TVSum) to form the F-Score for a given test video, and the way the average performance of these multiple runs is indicated (mean of highest performance for each run; best mean performance at the same training epoch for all runs)—has been adopted by the vast majority of the state-of-the-art works on video summarization (see [23], [25], [26], [30], [31], [33], [35], [37]–[40], [42], [45], [47], [48], [50]–[53], [96], [97], [127], [145], and [146]). Hence, it can be seen as the currently established approach for assessing the performance of video summarization algorithms.

A slightly different evaluation approach calculates the agreement with a single ground-truth summary, instead of multiple user summaries. This single ground-truth summary is computed by averaging the key-fragment summaries per frame, in the case of multiple key-fragment-based user summaries, and by averaging all users' scores for a video on a frame-basis, in the case of frame-level importance scores. This approach is utilized by only a few methods (i.e., [32], [34], [36], [49], and [54]), as it does not maintain the original opinion of every user, leading to a less firm evaluation.

Another evaluation approach was proposed in [147]. This method is independent of any predefined fragmentation of the video. The user-generated frame-level importance scores for the TVSum videos are considered as rankings, and two rank correlation coefficients, namely, Kendall τ [151] and Spearman ρ [152] coefficients, are used to evaluate the summary. However, these measures can be used only on datasets that follow the TVSum annotations and methods that produce the same type of results (i.e., frame-level importance scores). This methodology was used (in addition to the established protocol) to evaluate the algorithms in [44] and [48].

Last but not least, a new evaluation protocol for video summarization was presented in [148]. This work started by evaluating the performance of five publicly available methods under a large-scale experimental setting with 50 randomly created data splits of the SumMe and TVSum datasets, where the performance is evaluated using F-Score. The conducted study showed that the results reported in the relevant papers are not always congruent with the performance on the large-scale experiment, and the F-Score is not most suitable for comparing algorithms that were run on different splits. For this, Apostolidis *et al.* [148] proposed a new evaluation protocol, called "Performance over Random," which estimates the difficulty of each used data split and utilizes this information during the evaluation process.

V. PERFORMANCE COMPARISONS

A. Quantitative Comparisons

In this section, we present the performance of the reviewed deep-learning-based video summarization approaches on the SumMe and TVSum datasets, as reported in the corresponding papers. We focus only on these two datasets since they are prevalent in the relevant literature.

Table 4 reports the performance of (weakly) supervised video summarization algorithms that have been assessed via the established evaluation approach (i.e., using the entire set of available user summaries for a given video). In the same table, we report the performance of a random summarizer. To estimate this performance, importance scores are randomly assigned to the frames of a given video based on a uniform distribution of probabilities. The corresponding fragment-level scores are then used to form video summaries using the Knapsack algorithm and a length budget of a maximum of 15% of the original video's duration. Random summarization is performed 100 times for each video, and the overall average score is reported (for further details, please check the relevant algorithm in [148]). Moreover, the rightmost column of this table provides details about the number and type of data splits that were used for the evaluation, with "X Rand" denoting X randomly created splits [with X being equal to 1, 5, and 10 or an unspecified value (this case is noted as M Rand)] and "5 FCV" denoting fivefold cross validation. Finally, the algorithms are listed according to their average ranking on both datasets (see the fourth column "Avg Rnk").

Based on the reported performances, we can make the following observations.

1) The best-performing supervised approaches utilize tailored attention mechanisms (VASNet, H-MAN, SUM-GDA, DASP, and CSNet$_{sup}$) or memory

Table 4 Comparison [F1: F-Score (%)] of Supervised and Weakly Supervised Video Summarization Approaches on SumMe and TVSum. Weakly Supervised Methods Marked With ⋄. Multimodal Approaches Are Marked With +

	SumMe		TVSum		Avg	Data
	F1	Rnk	F1	Rnk	Rnk	splits
Random summary	40.2	27	54.4	24	25.5	–
vsLSTM [22]	37.6	30	54.2	25	27.5	1 Rand
dppLSTM [22]	38.6	29	54.7	23	26	1 Rand
⋄SASUM [127]	40.6	25	53.9	26	25.5	10 Rand
ActionRanking [31]	40.1	28	56.3	21	24.5	1 Rand
⋄FPVSF [109]	41.9	24	– [3]	–	24	–
vsLSTM+Att [24]	43.2	22	– [4]	–	22	1 Rand
H-RNN [23]	42.1	23	57.9	18	20.5	–
DR-DSN$_{sup}$ [40]	42.1	23	58.1	16	19.5	5 FCV
dppLSTM+Att [24]	43.8	19	– [4]	–	19	1 Rand
+DSSE [46]	–	–	57.0	19	19	–
⋄WS-HRL [44]	43.6	21	58.4	14	17.5	5 FCV
PCDL$_{sup}$ [42]	43.7	20	59.2	10	15	10 Rand
SF-CVS [103]	46.0	12	58.0	17	14.5	–
+SASUM$_{fullysup}$ [127]	45.3	14	58.2	15	14.5	10 Rand
UnpairedVSN$_{psup}$ [39]	48.0	7	56.1	22	14.5	5 Rand
SUM-FCN [28]	47.5	9	56.8	20	14.5	M Rand
MAVS [29]	40.3	26	**66.8**	1	13.5	5 FCV
A-AVS [26]	43.9	18	59.4	9	13.5	5 Rand
CRSum [101]	47.3	10	58.0	17	13.5	5 FCV
HSA-RNN [96]	44.1	17	59.8	8	12.5	–
+DQSN [45]	–	–	58.6	12	12	5 FCV
TTH-RNN [51]	44.3	16	60.2	7	11.5	–
M-AVS [26]	44.4	15	61.0	5	10	5 Rand
ACGAN$_{sup}$ [38]	47.2	11	59.4	9	10	5 FCV
SUM-DeepLab [28]	48.8	5	58.4	14	9.5	M Rand
CSNet$_{sup}$ [35]	48.6	6	58.5	13	9.5	5 FCV
DASP [27]	45.5	13	63.6	3	8	5 Rand
SMLD [53]	47.6	8	61.0	5	6.5	5 FCV
SUM-GDA [50]	52.8	2	58.9	11	6.5	5 FCV
H-MAN [97]	51.8	3	60.4	6	4.5	5 FCV
VASNet [25]	49.7	4	61.4	4	4	5 Rand
SMN [30]	**58.3**	1	64.5	2	**1.5**	1 Rand

networks (SMN) to capture variable- and long-range temporal dependencies.

2) Some works (e.g., MAVS, DASP, M-AVS, A-AVS, and TTH-RNN) exhibit high performance in one of the datasets and very low or even random performance in the other dataset. This poorly balanced performance indicates techniques that may be highly adapted to a specific dataset.
3) The use of data from additional modalities (such as text-based video metadata and descriptions) does not seem to help in the considered datasets, as the multimodals (DSSE, DQSN, and SASUM$_{\text{fullysup}}$) are not competitive compared to the unimodal ones that rely on the analysis of the visual content only.
4) The use of weak labels instead of a full set of human annotations does not enable a good summarization, as the weakly supervised methods perform poorly (SASUM, FPVSF, and WS-HRL).
5) Finally, a few methods (placed at the top of Table 4) show random performance in at least one of the used datasets.

[3]In this work, the TVSum dataset is used as the third-person labeled data for training purposes only, so we do not present any result here.

[4]The authors of this literature work evaluate their method on TVSum using a protocol that differs from the typical protocol used with this dataset, so we do not present this result here.

Table 5 Comparison [F1: F-Score (%)] of Unsupervised Video Summarization Approaches on SumMe and TVSum

	SumMe		TVSum		Avg	Data
	F1	Rnk	F1	Rnk	Rnk	splits
Random summary	40.2	13	54.4	11	12	–
Online Motion-AE [107]	37.7	14	51.5	13	13.5	–
SUM-FCN$_{unsup}$ [28]	41.5	11	52.7	12	11.5	M Rand
DR-DSN [40]	41.4	12	57.6	8	10	5 FCV
EDSN [41]	42.6	10	57.3	9	9.5	5 FCV
UnpairedVSN [39]	47.5	7	55.6	10	8.5	5 Rand
PCDL [42]	42.7	9	58.4	6	7.5	10 Rand
ACGAN [38]	46.0	8	58.5	5	6.5	5 FCV
SUM-GAN-sl [33]	47.8	6	58.4	6	6	5 Rand
SUM-GAN-AAE [37]	48.9	5	58.3	7	6	5 Rand
SUM-GDA$_{unsup}$ [50]	50.0	4	59.6	2	3	5 FCV
CSNet+GL+RPE [48]	50.2	3	59.1	3	3	5 FCV
CSNet [35]	**51.3**	1	58.8	4	2.5	5 FCV
AC-SUM-GAN [47]	50.8	2	**60.6**	1	**1.5**	5 Rand

Table 5 presents the performance of unsupervised video summarization methods that have been assessed with the same evaluation approach (i.e., using the entire set of available user summaries for a given video). As in Table 4, Table 5 also reports the performance of a random summarizer and provides details about the number and type of data splits that were used for the evaluation (see the rightmost column). Once again, the algorithms are presented according to their average ranking on both datasets (see the fourth column "Avg Rnk").

Based on the reported results, we can make the following remarks.

1) The use of GANs for learning summarization in a fully unsupervised manner is a good choice, as most of the best-performing methods (AC-SUM-GAN, CSNet, CSNet+GL+RPE, SUM-GAN-AAE, and SUM-GAN-sl) rely on this framework.
2) The use of attention mechanisms helps to identify the important parts of the video, as a few of the best-performing algorithms (CSNet, CSNet+GL+RPE, SUM-GDA$_{\text{unsup}}$, and SUM-GAN-AAE) utilize such mechanisms. The benefits of using such a mechanism are also documented through the comparison of the SUM-GAN-sl and SUM-GAN-AAE techniques. The replacement of the VAE (that is used in SUM-GAN-sl) by a deterministic attention autoencoder (which is introduced in SUM-GAN-AAE) results in a clear performance improvement on the SumMe dataset while maintaining the same levels of summarization performance on TVSum.
3) Techniques that rely on reward functions and reinforcement learning (DR-DSN, EDSN) are not so competitive compared to GAN-based methods, especially on SumMe.
4) Finally, a few methods (placed at the top of Table 5) perform approximately equally to the random summarizer.

In Table 6, we show the performance of video summarization methods that have been evaluated with a variation of the established protocol, i.e., by comparing each generated summary with a single ground-truth

Table 6 Comparison [F1: F-Score (%)] of Video Summarization Approaches on SumMe and TVSum, Using a Single Ground-Truth Summary for Each Video. Unsupervised Methods Are Marked With ⋆

	SumMe		TVSum		Avg	Data
	F1	Rnk	F1	Rnk	Rnk	splits
Random Summary	40.2	8	54.4	9	8.5	–
⋆SUM-GAN [32]	38.7	10	50.8	11	10.5	5 Rand
⋆SUM-GAN$_{dpp}$ [32]	39.1	9	51.7	10	9.5	5 Rand
SUM-GAN$_{sup}$ [32]	41.7	7	56.3	8	7.5	5 Rand
⋆Cycle-SUM [36]	41.9	6	57.6	7	6.5	5 Rand
DTR-GAN [54]	–	–	61.3	6	6	1 Rand
Ptr-Net [34]	46.2	5	63.6	4	4.5	–
⋆SUM-Ind_{LU} [49]	51.4	3	61.5	5	4	5 FCV
⋆SUM-GAN-sl [33]	46.8	4	**65.3**	1	2.5	5 Rand
⋆SUM-GAN-AAE [37]	56.9	2	63.9	3	2.5	5 Rand
⋆AC-SUM-GAN [47]	**60.7**	1	64.8	2	**1.5**	5 Rand

summary per video (see Section IV-B2, third paragraph). As before, this table reports the performance of a random summarizer according to this evaluation approach, provides details about the number and type of data splits that were used for the evaluation (see the rightmost column), and lists the algorithms according to their average ranking on both datasets (see the fourth column "Avg Rnk"). Our remarks on the reported data are given as follows.

1) Only a limited number of video summarization works rely on the use of the single ground-truth summary for evaluating the summarization performance.
2) Among the supervised methods, Ptr-Net is the top-performing one in both datasets. However, from the reportings in the relevant paper, it is not clear how many different randomly created data splits were used for evaluation.
3) Concerning the unsupervised approaches (marked with an asterisk), the SUM-GAN method and its extension that uses the DPP to increase the diversity of the summary content (called SUM-GAN$_{dpp}$) perform worse than a random summarizer. However, three newer extensions of this general approach that was evaluated also with this protocol (i.e., in addition to assessments made using the established evaluation approach), namely, the SUM-GAN-sl, SUM-GAN-AAE, and AC-SUM-GAN methods, exhibit very good performance that even surpasses the performance of the (few) supervised approaches listed in this table.

Having discussed the results reported in each of Tables 4–6 alone, at this point, we extend our observations by some additional remarks. The F-Score values reported in Tables 4 and 5 show that the use of a supervision signal (associated with the ground-truth data) to train a method originally designed as an unsupervised one (e.g., DR-DSN, PCDL, ACGAN, and CSNet) does not lead to considerably improved performance. Hence, focusing on purely supervised or unsupervised methods and trying to explore their learning capacity seem to be a more effective approach. Furthermore, purely unsupervised methods can be competitive to supervised ones (e.g., AC-SUM-GAN and CSNet), and thus, additional future efforts toward the improvement of such algorithms are definitely in the right direction.

With regards to the evaluation of video summarization algorithms, the rightmost column in all these three tables clearly indicates a lack of consistency. Most algorithms evaluate the summarization performance on one, five, or ten randomly created data splits, while, in some papers, this information is completely missing (we denote this case by "M Rand" in these tables). Moreover, randomly created data splits may exhibit some overlap among them, a case that differs from the fivefold cross validation (see "5 FCV" in the tables) approach that is adopted by fewer methods. This observed diversity in the implementation of the employed evaluation protocol, along with the use of different randomly created splits in most papers (see relevant considerations in [148]), unfortunately, does not allow for a perfectly accurate performance comparison between the different summarization algorithms.

Last but not least, another open issue of the relevant bibliography relates to a lack of information with respect to the applied approach for terminating the training process and selecting the trained model. Concerning the (weakly) supervised methods, a few of them (e.g., [22], [28]–[31], [45], and [101]) explicitly state that model selection relies on the summarization performance on a validation set. One of them [46] terminates the training process based on a predefined condition (training terminates if the average training loss difference between two consecutive epochs is less than a threshold that relies on the initial loss value). Most of them (e.g., [23], [34], [42], [51], [53], [54], [97], [96], [103], and [127]) do not provide information regarding the use of a validation set or the application of a criterion for terminating the training process. Regarding the unsupervised approaches, one of them that relies on reinforcement learning [40] states that training terminates upon a condition that relates to the received reward; i.e., training stops after reaching a maximum number of epochs (60 epochs), while early stopping is executed when the received reward stops to increase for a particular time period (10 epochs). Another recent approach [47], which integrates an actor–critic model into a GAN and uses the discriminator's feedback as a reward signal, selects a well-trained model based on a criterion that maximizes the overall received reward and minimizes the Actor's loss. A couple of methods [35], [41] end the training process based on a maximum number of training epochs and then select the last trained model. Once again, most works (e.g., [32], [33], [36]–[39], [48], and [107]) do not provide details about the use of a termination criterion. Nevertheless, experimentation with a few methods with publicly available implementations (i.e., [25], [33], [37], and [40]) showed that the learning/performance curve can exhibit fluctuations; in Fig. 8, the learning/performance curve of the examined supervised (VASNet) and unsupervised (DR-DSN, SUM-GAN-sl, and SUM-GAN-AAE) algorithms shows noticeable fluctuations even after a good number of training epochs. This fluctuation denotes that the networks, in general, are able to develop knowledge about the

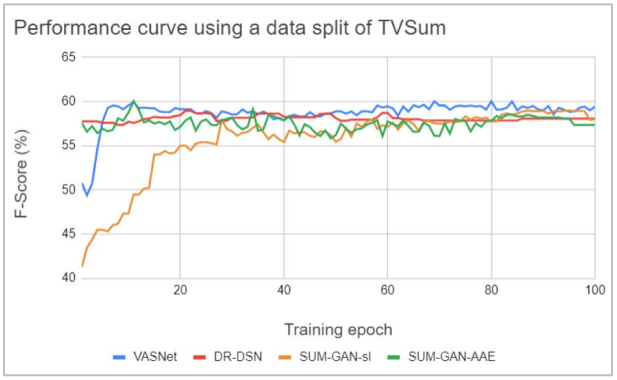


Fig. 8. ***Performance of four summarization methods on a set of test videos of the TVSum dataset as the training proceeds.***

task (sometimes even in the very early training epochs), but the selection of the best-trained model at the end is not always straightforward.

B. Qualitative Comparisons and Demos

In addition to the numerical comparisons discussed above, and with a view to gaining an intuitive understanding of the summarization results, in the sequel, we present the summaries produced for video #19 of SumMe ("St Maarten Landing") by five publicly available summarization methods. The frame sequence at the top of Fig. 9 represents the flow of the story depicted in the video. Below this brief overview, for each considered method, there is a diagram that shows the selected shots of the video and a visualization of the most important shots by a representative key frame for each. Specifically, the gray bars represent the average human-annotated importance scores, the black vertical lines show the boundaries of each shot, and the colored bars are the shots that each method has selected for inclusion in the summary. We observe that SUM-GAN-AAE and VASNet manage to select the shots with the highest importance, and for this, they also achieve the highest F-Score. In addition, the key frames they select are diverse and give a good overview of the plane's landing. DR-DSN selects almost the same shots as the two aforementioned methods, leading to a bit lower performance. SUM-GAN-sl focuses a bit less on the main event of the video; dppLSTM loses the point of the video and selects many frames that show only the background without the plane, leading to a poor F-Score.

Furthermore, to get an idea of how such summarization methods work in practice, one can experiment with tools such as the "On-line Video Summarization Service"[5] of [153], which integrates an adaptation of the SUM-GAN-AAE algorithm [37]. This tool enables the creation of multiple summaries for a given video, which are tailored to the needs of different target distribution channels (i.e., different video sharing/social networking platforms).

[5]http://multimedia2.iti.gr/videosummarization/service/start.html

VI. FUTURE DIRECTIONS

Given the current state of the art in automated video summarization, we argue that future work in this field should primarily targets the development of deep learning methods that can be trained effectively without the need for large collections of human-annotated ground-truth data. In this way, the research community will be able to tackle issues associated with the limited amount of annotated data and significantly diminish (or even completely eliminate) the need for laborious and time-demanding data annotation tasks. In this direction, the research community should put efforts toward the design and development of deep learning architectures that can be trained either in a fully unsupervised or in a semisupervised/weakly supervised manner.

A. Unsupervised Learning

With respect to the development of unsupervised video summarization methods, given the fact that most of the existing approaches try to increase the representativeness of the generated summary with the help of summary-to-video reconstruction mechanisms, future work could investigate mechanisms that force the outcome of the summarization process to meet additional criteria about the content of the generated summary, such as its visual diversity (that was considered in [39], [40], [50], and [95]) and its uniformity (that was examined in [49]). On a similar basis, efforts could focus on extending current deep learning architectures that combine the merits of adversarial and reinforcement learning [47], by utilizing a soft actor–critic [154] that is capable of further discovering the action space via automatically defining a suitable value for the entropy regularization factor, and by introducing additional rewards that relate to the additional summarization criteria, such as the aforementioned ones.

B. Weakly Supervised Learning

With regards to the development of semisupervised or weakly supervised approaches, the goal would be to investigate ways to intervene in the summary production process so that the outcome (i.e., a video summary) is aligned with user-specified rules. One approach in this direction is the generation of a summary according to a set of textual queries that indicate the desired summary content (as in [155]–[159]). Another, the more aspiring approach would be the use of an online interaction channel between the user/editor and the trainable summarizer, in combination with active learning algorithms that allow to incorporate the user's/editor's feedback with respect to the generated summary (as in [160]). Finally, the possibility of adapting graph signal processing approaches [161], which have already been applied with success to data sampling [162] and image/video analysis tasks [163], [164], for introducing such external supervision could be examined. The development of effective semisupervised or weakly supervised summarization approaches will allow to

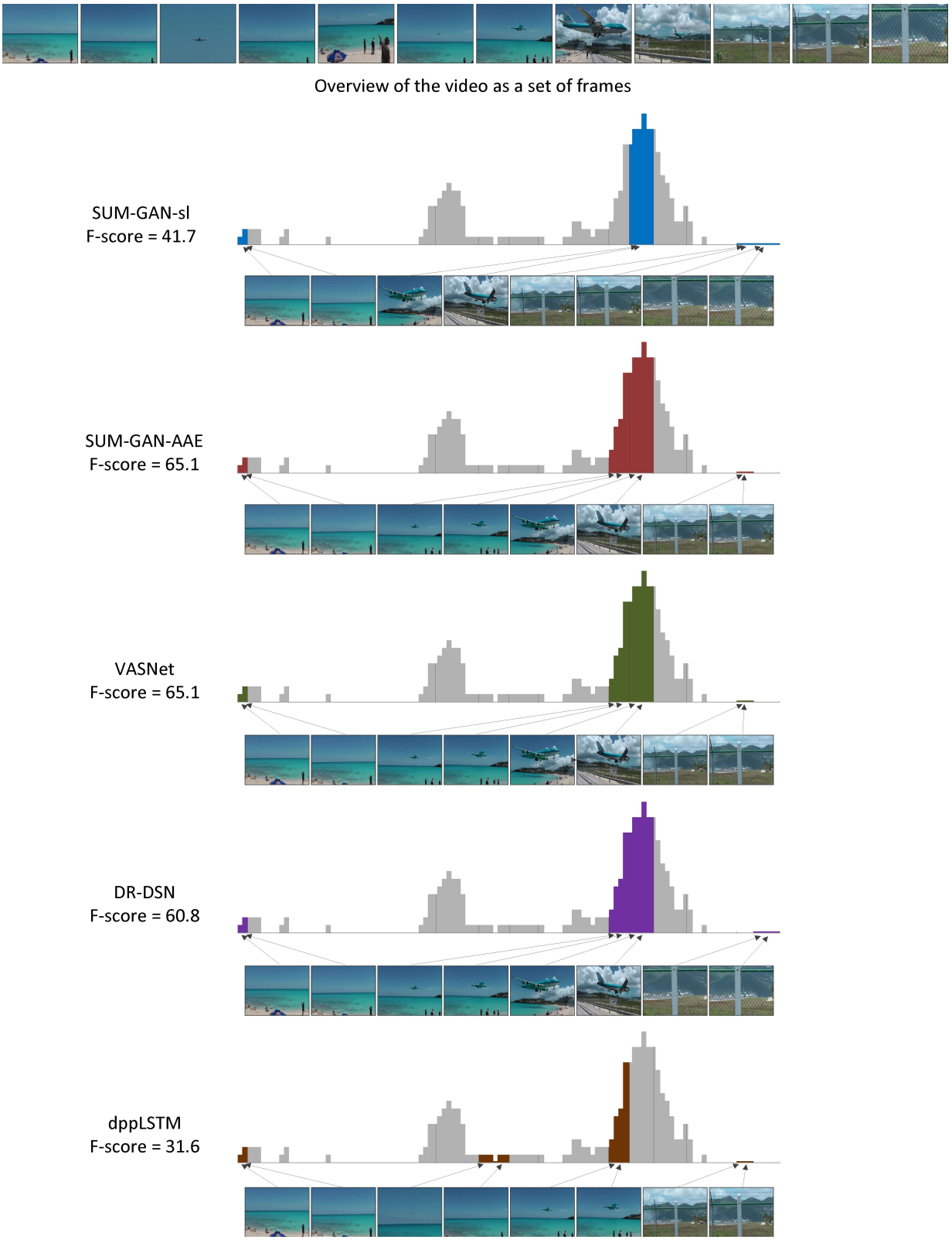


Fig. 9. ***Overview of video #19 of the SumMe dataset and the produced summaries by five summarization algorithms with publicly available implementations.***

better meet the needs of specific summarization scenarios and application domains. For example, such developments are often important for the practical application of summarization technologies in the News/Media Industry, where complete automation that diminishes editorial control over the generated summaries is not always preferred.

C. Generative Adversarial Networks

Concerning the training of unsupervised video summarization methods, we show that most of these methods rely on the adversarial training of GANs. However, open questions with respect to the training of such architectures, such as sufficient convergence conditions and mode collapse, still remain. Thus, another promising research direction could be to investigate ways to improve the training process. For this, one strategy could be the use of augmented training data (that do not require human annotation) in combination with curriculum learning approaches. Such approaches have already been examined for improving the training of GANs (see [165]–[167]) in applications other than video summarization. We argue that transferring the gained knowledge from these works to the video summarization domain would contribute to advancing the effectiveness of unsupervised GAN-based summarization approaches. Regarding the training of semisupervised or weakly supervised video summarization methods, besides the use of an online interaction channel between the user/editor and the trainable summarizer that was discussed in the previous paragraph, supervision could also relate to the use of an adequately large set of unpaired data (i.e., raw videos and video summaries with no correspondence between them) from a particular summarization domain or application scenario, as in [39]. Such a data-driven weak-supervision approach could possibly eliminate the need for fine-grained supervision signals (i.e., human-generated ground-truth annotations for the collection of the raw videos) or handcrafted functions that model the domain rules (which, in most cases, are really hard to obtain), and would allow a deep learning architecture to automatically learn a mapping function between the raw videos and the summaries in the targeted domain.

D. Recurrent Neural Networks

Another future research objective involves efforts to overcome the identified weaknesses of using RNNs for video summarization that was discussed in e.g., [25] and [49]–[51] and mainly relate to the computationally demanding and hard-to-parallelize training process, as well as to the limited memory capacity of these networks. For this, future work could examine the use of IndRNNs [105] that were shown to alleviate the drawbacks of LSTMs with respect to decaying, vanishing, and exploding gradients [49], in combination with high-capacity memory networks, such as the ones used in [29] and [30]. Alternatively, future work could build on existing approaches [25], [48], [50], [103] and develop more advanced attention mechanisms that encode the relative position of video frames and model their temporal dependencies according to different granularities (e.g., considering the entire frame sequence or also focusing on smaller parts of it). Such methods would be particularly suited for summarizing long videos (e.g., movies). Finally, with respect to video content representation, the above-proposed research directions could also involve the use of network architectures that model the spatiotemporal structure of the video, such as 3-D-CNNs and convolutional LSTMs.

E. Multimodal Data

With respect to the utilized data modality for learning a summarizer, currently, the focus is on the visual modality. Nevertheless, the audio modality of the video could be a rich source of information as well. For example, the audio content could help to automatically identify the most thrilling parts of a movie that should appear in a movie trailer. Moreover, the temporal segmentation of the video based also on the audio stream could allow the production of summaries that offer a more natural story narration compared to the generated summaries based on approaches that rely solely on the visual stream. We argue that deep learning architectures that have been utilized to model frames' dependencies based on their visual content could be examined also for analyzing the audio modality. In the following, the extracted representations from these two modalities could be fused according to different strategies (e.g., after exploring the latent consistency between them), to better indicate the most suitable parts for inclusion in the video summary.

F. Benchmarking

Finally, besides the aforementioned research directions that relate to the development and training of deep-learning-based architectures for video summarization, we strongly believe that efforts should be put toward the definition of better evaluation protocols to allow accurate comparison of the developed methods in the future. The discussions in [147] and [148] showed that the existing protocols have some imperfections that affect the reliability of performance comparisons. To eliminate the impact of the choices made when evaluating a summarization algorithm (that, e.g., relate to the split of the utilized data or the number of different runs), the relevant community should consider all the different parameters of the evaluation pipeline and precisely define a protocol that leaves no questions about the experimental outcomes of a summarization work. Then, the adoption of this protocol by the relevant community will enable fair and accurate performance comparisons.

VII. CONCLUSIONS

In this work, we provided a systematic review of the deep-learning-based video summarization landscape. This review allowed us to discuss how the summarization technology has evolved over the last years and what is the potential for the future, as well as to raise awareness to the relevant community with respect to promising future directions and open issues. The main conclusions of this study are outlined in the following.

Concerning the summarization performance, the best-performing supervised methods thus far learn frames' importance by modeling the variable-range temporal dependence among video frames/fragments with the help of recurrent neural networks and tailored attention mechanisms. The extension of the memorization capacity of LSTMs by using memory networks has shown promising results and should be further investigated. In the direction of unsupervised video summarization, the use of GANs for learning how to build a representative video summary seems to be the most promising approach. Such networks have been integrated into summarization architectures and used in combination with attention mechanisms or actor–critic models, showing a summarization performance that is comparable to the performance of state-of-the-art supervised approaches. Given the objective difficulty to create large-scale datasets with human annotations for training summarization models in a supervised way, further research effort should be put on the development of fully unsupervised or semisupervised/weakly supervised video summarization methods that eliminate or reduce to a large extent the need for such data and facilitate adaptation to the summarization requirements of different domains and application scenarios.

Regarding the evaluation of video summarization algorithms, there is some diversity among the used evaluation protocols in the bibliography, which is associated with the way that the used data are being divided for training and testing purposes, the number of the conducted experiments using different randomly created splits of the data, and the used data splits; concerning the latter, a recent work [148] showed that different randomly created data splits of the SumMe and TVSum datasets are characterized by considerably different levels of difficulty. All the above raise concerns regarding the accuracy of performance comparisons that rely on the results reported in the different papers. Moreover, there is lack of information in the reportings of several summarization works, with respect to the applied process for terminating the training process and selecting the trained model. Hence, the relevant community should be aware of these issues and take the necessary actions to increase the reproducibility of the results reported for each newly proposed method.

Last but not least, this work indicated several research directions toward further advancing the performance of video summarization algorithms. Besides these proposals for future scientific work, we believe that further efforts should be put toward the practical use of summarization algorithms, by integrating such technologies into tools that support the needs of modern media organizations for time-efficient video content adaptation and reuse. ■

REFERENCES

[1] M. Barbieri, L. Agnihotri, and N. Dimitrova, "Video summarization: Methods and landscape," in *Internet Multimedia Management Systems IV*, vol. 5242, J. R. Smith, S. Panchanathan, and T. Zhang, Eds. Bellingham, WA, USA: SPIE, 2003, pp. 1–13.

[2] Y. Li, S. Lee, C.-H. Yeh, and C.-C. J. Kuo, "Techniques for movie content analysis and skimming: Tutorial and overview on video abstraction techniques," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 79–89, Mar. 2006.

[3] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 3, no. 1, p. 3, Feb. 2007.

[4] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *J. Vis. Commun. Image Represent.*, vol. 19, no. 2, pp. 121–143, Feb. 2008.

[5] R. M. Jiang, A. H. Sadka, and D. Crookes, *Advances in Video Summarization and Skimming*. Berlin, Germany: Springer, 2009, pp. 27–50.

[6] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 797–819, Nov. 2011.

[7] M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah, "Video summarization: Techniques and classification," in *Computer Vision and Graphics*, L. Bolc, R. Tadeusiewicz, L. J. Chmielewski, and K. Wojciechowski, Eds. Berlin, Germany: Springer, 2012, pp. 1–13.

[8] A. G. del Molino, C. Tan, J.-H. Lim, and A.-H. Tan, "Summarization of egocentric videos: A comprehensive survey," *IEEE Trans. Hum.-Machine Syst.*, vol. 47, no. 1, pp. 65–76, May 2017.

[9] M. Basavarajaiah and P. Sharma, "Survey of compressed domain video summarization techniques," *ACM Comput. Surveys*, vol. 52, no. 6, pp. 1–29, Jan. 2020.

[10] V. V. K., D. Sen, and B. Raman, "Video skimming: Taxonomy and comprehensive survey," *ACM Comput. Surveys*, vol. 52, no. 5, pp. 1–38, Oct. 2019.

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

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

[13] M. Ma, S. Mei, S. Wan, Z. Wang, D. D. Feng, and M. Bennamoun, "Similarity based block sparse subset selection for video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3967–3980, Oct. 2021.

[14] Y. Li, T. Zhang, and D. Tretter, "An overview of video abstraction techniques," Hewlett Packard, Palo Alto, CA, USA, Tech. Rep. HPL-2001-191, Jan. 2001.

[15] J. Calic, D. P. Gibson, and N. W. Campbell, "Efficient layout of comic-like video summaries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 7, pp. 931–936, Jul. 2007.

[16] T. Wang, T. Mei, X.-S. Hua, X.-L. Liu, and H.-Q. Zhou, "Video collage: A novel presentation of video sequence," in *Proc. IEEE Multimedia Expo Int. Conf.*, Jul. 2007, pp. 1479–1482.

[17] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 1–9.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[22] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 766–782.

[23] B. Zhao, X. Li, and X. Lu, "Hierarchical recurrent neural network for video summarization," in *Proc. 25th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2017, pp. 863–871.

[24] L. L. Casas and E. Koblents, "Video summarization with LSTM and deep attention models," in *MultiMedia Modeling*, I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W.-H. Cheng, and S. Vrochidis, Eds. Cham, Switzerland: Springer, 2019, pp. 67–79.

[25] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Proc. Asian Conf. Comput. Vis. (ACCV) Workshops*, G. Carneiro and S. You, Eds. Cham, Switzerland: Springer, 2019, pp. 39–54.

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

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

[28] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 358–374.

[29] L. Feng, Z. Li, Z. Kuang, and W. Zhang, "Extractive video summarizer with memory augmented

neural networks," in *Proc. 26th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2018, pp. 976–983.

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

[31] M. Elfeki and A. Borji, "Video summarization via actionness ranking," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 754–763.

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

[33] E. Apostolidis, A. I. Metsai, E. Adamantidou, V. Mezaris, and I. Patras, "A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization," in *Proc. 1st Int. Workshop AI Smart TV Content Prod., Access Del. (AI TV)*, New York, NY, USA, 2019, pp. 17–25.

[34] T.-J. Fu, S.-H. Tai, and H.-T. Chen, "Attentive and adversarial learning for video summarization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 1579–1587.

[35] Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon, "Discriminative feature learning for unsupervised video summarization," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8537–8544.

[36] L. Yuan, F. E. H. Tay, P. Li, L. Zhou, and J. Feng, "Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9143–9150.

[37] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Unsupervised video summarization via attention-driven adversarial learning," in *Proc. 26th Int. Conf. Multimedia Modeling (MMM)*. Cham, Switzerland: Springer, 2020, pp. 492–504.

[38] X. He *et al.*, "Unsupervised video summarization with attentive conditional generative adversarial networks," in *Proc. 27th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2019, pp. 2296–2304.

[39] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7894–7903.

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

[41] N. Gonuguntla *et al.*, "Enhanced deep video summarization network," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019.

[42] B. Zhao, X. Li, and X. Lu, "Property-constrained dual learning for video summarization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3989–4000, Oct. 2020.

[43] S. Cai, W. Zuo, L. S. Davis, and L. Zhang, "Weakly-supervised video summarization using variational encoder-decoder and web prior," in *Proc. Euro. Conf. Comput. Vis. (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 193–210.

[44] Y. Chen, L. Tao, X. Wang, and T. Yamasaki, "Weakly supervised video summarization by hierarchical reinforcement learning," in *Proc. ACM Multimedia Asia (MMAsia)*, Beijing, China. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–6, Art. no. 3, doi: 10.1145/3338533.3366583.

[45] K. Zhou, T. Xiang, and A. Cavallaro, "Video summarisation by classification with deep reinforcement learning," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018.

[46] Y. Yuan, T. Mei, P. Cui, and W. Zhu, "Video summarization by learning deep side semantic embedding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 226–237, Jan. 2019.

[47] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3278–3292, Aug. 2021.

[48] Y. Jung, D. Cho, S. Woo, and I. S. Kweon, "Global-and-local relative position embedding for unsupervised video summarization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 167–183.

[49] G. Yaliniz and N. Ikizler-Cinbis, "Using independently recurrent networks for reinforcement learning based unsupervised video summarization," *Multimedia Tools Appl.*, vol. 80, no. 12, pp. 17827–17847, May 2021, doi: 10.1007/s11042-020-10293-x.

[50] P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, "Exploring global diverse attention via pairwise temporal relation for video summarization," *Pattern Recognit.*, vol. 111, Mar. 2021, Art. no. 107677. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320304805

[51] B. Zhao, X. Li, and X. Lu, "TTH-RNN: Tensor-train hierarchical recurrent neural network for video summarization," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3629–3637, Apr. 2020.

[52] S. Lal, S. Duggal, and I. Sreedevi, "Online video summarization: Predicting future to better summarize present," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 471–480.

[53] W.-T. Chu and Y.-H. Liu, "Spatiotemporal modeling and label distribution learning for video summarization," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2019, pp. 1–6.

[54] Y. Zhang, M. Kampffmeyer, X. Zhao, and M. Tan, "DTR-GAN: Dilated temporal relational adversarial network for video summarization," in *Proc. ACM Turing Celebration Conf. (ACM TURC)*, New York, NY, USA, May 2019, p. 89.

[55] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 505–520. [Online]. Available: https://gyglim.github.io/me/

[56] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *Proc. CVPR*, Jun. 2015, pp. 5179–5187. [Online]. Available: https://github.com/yalesong/tvsum

[57] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006, doi: 10.1162/neco.2006.18.7.1527.

[58] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted Boltzmann machines for collaborative filtering," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2007, pp. 791–798, doi: 10.1145/1273496.1273596.

[59] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proc. 12th Int. Conf. Artif. Intell. Statist.*, vol. 5, D. van Dyk and M. Welling, Eds. Clearwater Beach, FL, USA: Hilton Clearwater Beach Resort, Apr. 2009, pp. 448–455. [Online]. Available: http://proceedings.mlr.press/v5/salakhutdinov09a.html

[60] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," in *Proc. 6th Int. Conf. Neural Inf. Process. Syst.* San Francisco, CA, USA: Morgan Kaufmann, 1993, pp. 3–10.

[61] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014.

[62] Y. LeCun and Y. Bengio, *Convolutional Networks for Images, Speech, and Time Series*. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258.

[63] C. Goller and A. Kuchler, "Learning task-dependent distributed representations by backpropagation through structure," in *Proc. IEEE Int. Conf. Neural Netw.*, vol. 1, Jun. 1996, pp. 347–352.

[64] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Representations by Back-Propagating Errors*. Cambridge, MA, USA: MIT Press, 1988, pp. 696–699.

[65] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[66] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[67] J. Gast and S. Roth, "Lightweight probabilistic deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3369–3378. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Gast_Lightweight_Probabilistic_Deep_CVPR_2018_paper.html

[68] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231216315533

[69] S. Pouyanfar *et al.*, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–36, Sep. 2018, doi: 10.1145/3234150.

[70] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100379. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013721000198

[71] H. Schwenk, "Continuous space translation models for phrase-based statistical machine translation," in *Proc. COLING, Posters*. Mumbai, India: The COLING Organizing Committee, Dec. 2012, pp. 1071–1080. [Online]. Available: https://www.aclweb.org/anthology/C12-2104

[72] L. Dong, F. Wei, K. Xu, S. Liu, and M. Zhou, "Adaptive multi-compositionality for recursive neural network models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 422–431, Mar. 2016, doi: 10.1109/TASLP.2015.2509257.

[73] Y. You, Y. Qian, T. He, and K. Yu, "An investigation on DNN-derived bottleneck features for GMM-HMM based robust speech recognition," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Jul. 2015, pp. 30–34.

[74] A. L. Maas *et al.*, "Building DNN acoustic models for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 41, pp. 195–213, Jan. 2017, doi: 10.1016/j.csl.2016.06.007.

[75] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1196–1206, May 2016.

[76] 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*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham, Switzerland: Springer, 2020, pp. 483–492.

[77] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[78] H. Xiao, J. Feng, G. Lin, Y. Liu, and M. Zhang, "MoNet: Deep motion exploitation for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1140–1148.

[79] P. Xu, M. Ye, X. Li, Q. Liu, Y. Yang, and J. Ding, "Dynamic background learning through deep auto-encoder networks," in *Proc. 22nd ACM Int. Conf. Multimedia*, New York, NY, USA, Nov. 2014, pp. 107–116, doi: 10.1145/2647868.2654914.

[80] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320317304120

[81] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *Neural Netw.*, vol. 117, pp. 8–66, Sep. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608019301303

[82] P. Dixit and S. Silakari, "Deep learning algorithms for cybersecurity applications: A technological and status review," *Comput. Sci. Rev.*, vol. 39, Feb. 2021, Art. no. 100317. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013720304172

[83] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A survey on deep learning for neuroimaging-based brain disorder analysis," *Frontiers Neurosci.*, vol. 14, p. 779, Oct. 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2020.00779

[84] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Comput.*, vol. 32, no. 5, pp. 829–864, May 2020.

[85] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 17, 2020, doi: 10.1109/TKDE.2020.2981314.

[86] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, Apr. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21918

[87] K. K. Thekumparampil, A. Khetan, Z. Lin, and S. Oh, "Robustness of conditional gans to noisy labels," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2018, pp. 10292–10303.

[88] A. Creswell and A. A. Bharath, "Denoising adversarial autoencoders," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 968–984, Apr. 2019.

[89] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in Neural Information Processing Systems*, vol. 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf

[90] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," 2018, *arXiv:1610.01644*. [Online]. Available: https://arxiv.org/abs/1610.01644

[91] M. Aubry and B. C. Russell, "Understanding deep features with computer-generated imagery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2875–2883.

[92] C. Yun, S. Sra, and A. Jadbabaie, "A critical view of global optimality in deep learning," in *Proc. Int. Conf. Mach. Learn. Represent.*, 2018.

[93] Z. Zheng and P. Hong, "Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2018, pp. 7924–7933.

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

[95] A. Kulesza and B. Taskar, *Determinantal Point Processes for Machine Learning*. Hanover, MA, USA: Now, 2012.

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

[97] Y.-T. Liu, Y.-J. Li, F.-E. Yang, S.-F. Chen, and Y.-C.-F. Wang, "Learning hierarchical self-attention for video summarization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3377–3381.

[98] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[99] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[100] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[101] Y. Yuan, H. Li, and Q. Wang, "Spatiotemporal modeling for video summarization using convolutional recurrent neural network," *IEEE Access*, vol. 7, pp. 64676–64685, 2019.

[102] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Doha, Qatar: ACL, Oct. 2014, pp. 1724–1734.

[103] C. Huang and H. Wang, "A novel key-frames selection framework for comprehensive video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 577–589, Feb. 2020.

[104] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 2692–2700.

[105] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (IndRNN): Building a longer and deeper RNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5457–5466.

[106] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 20–36.

[107] Y. Zhang, X. Liang, D. Zhang, M. Tan, and E. P. Xing, "Unsupervised object-level video summarization with online motion auto-encoder," *Pattern Recognit. Lett.*, vol. 130, pp. 376–385, Feb. 2020.

[108] R. Panda, A. Das, Z. Wu, J. Ernst, and A. K. Roy-Chowdhury, "Weakly supervised summarization of web videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3677–3686.

[109] H.-I. Ho, W.-C. Chiu, and Y.-C. F. Wang, "Summarizing first-person videos from third persons' points of views," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 72–89. [Online]. Available: https://github.com/azuxmioy/fpvsum

[110] R. Ren, H. Misra, and J. M. Jose, "Semantic based adaptive movie summarisation," in *Proc. 16th Int. Conf. Adv. Multimedia Modeling (MMM)*. Berlin, Germany: Springer-Verlag, 2010, pp. 389–399.

[111] F. Wang and C.-W. Ngo, "Summarizing rushes videos by motion, object, and event understanding," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 76–87, Feb. 2012.

[112] V. Kiani and H. R. Pourreza, "Flexible soccer video summarization in compressed domain," in *Proc. ICCKE*, Oct. 2013, pp. 213–218.

[113] Y. Li, B. Merialdo, M. Rouvier, and G. Linares, "Static and dynamic video summaries," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2011, pp. 1573–1576.

[114] C. Li, Y. Xie, X. Luan, K. Zhang, and L. Bai, "Automatic movie summarization based on the visual-audio features," in *Proc. IEEE 17th Int. Conf. Comput. Sci. Eng.*, Dec. 2014, pp. 1758–1761.

[115] G. Evangelopoulos *et al.*, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

[116] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, and A. Potamianos, "Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 4361–4365.

[117] B. A. Plummer, M. Brown, and S. Lazebnik, "Enhancing video summarization via vision-language embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1052–1060.

[118] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Multi-modal summarization for asynchronous collection of text, image, audio and video," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Copenhagen, Denmark, Sep. 2017, pp. 1092–1102.

[119] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for How2 videos," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6587–6596.

[120] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for How2 videos," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 6587–6596.

[121] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, and C. Zong, "MSMO: Multimodal summarization with multimodal output," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 4154–4164.

[122] M. Sanabria, Sherly, F. Precioso, and T. Menguy, "A deep architecture for multimodal summarization of soccer games," in *Proc. 2nd Int. Workshop Multimedia Content Anal. Sports (MMSports)*, New York, NY, USA, 2019, pp. 16–24.

[123] Y. Li, A. Kanemura, H. Asoh, T. Miyanishi, and M. Kawanabe, "Extracting key frames from first-person videos in the common space of multiple sensors," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3993–3997.

[124] X. Song *et al.*, "Category driven deep recurrent neural network for video summarization," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2016, pp. 1–6.

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

[126] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Video summarization using deep semantic features," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham, Switzerland: Springer, 2017, pp. 361–377.

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

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

[129] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3584–3592. [Online]. Available: https://github.com/l2ior/cosum

[130] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 540–555. [Online]. Available: http://lear.inrialpes.fr/people/potapov/med_summaries

[131] K.-H. Zeng, T.-H. Chen, J. C. Niebles, and M. Sun, "Title generation for user generated videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham,

Switzerland: Springer, 2016, pp. 609–625. [Online]. Available: http://aliensunmin.github.io/project/video-language/

[132] C.-Y. Fu, J. Lee, M. Bansal, and A. Berg, "Video highlight prediction using audience chat reactions," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 972–978.

[133] S. E. F. de Avila, A. D. Jr., A. de A. Araújo, and M. Cord, "VSUMM: An approach for automatic video summarization and quantitative evaluation," in *Proc. 21st Brazilian Symp. Comput. Graph. Image Process.*, Oct. 2008, pp. 103–110.

[134] N. Ejaz, I. Mehmood, and S. W. Baik, "Feature aggregation based visual attention model for video summarization," *Comput. Elect. Eng.*, vol. 40, no. 3, pp. 993–1005, Apr. 2014.

[135] V. Chasanis, A. Likas, and N. Galatsanos, "Efficient video shot summarization using an enhanced spectral clustering approach," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, V. Kůrková, R. Neruda, and J. Koutník, Eds. Berlin, Germany: Springer, 2008, pp. 847–856.

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

[137] J. Almeida, N. J. Leite, and R. D. S. Torres, "VISON: VIdeo Summarization for ONline applications," *Pattern Recognit. Lett.*, vol. 33, no. 4, pp. 397–409, Mar. 2012.

[138] E. J. Y. C. Cahuina and G. C. Chavez, "A new method for static video summarization using local descriptors and video temporal segmentation," in *Proc. 26th Conf. Graph., Patterns Images*, Aug. 2013, pp. 226–233.

[139] H. Jacob, F. L. C. Padua, A. Lacerda, and A. C. M. Pereira, "A video summarization approach based on the emulation of bottom-up mechanisms of visual attention," *J. Intell. Inf. Syst.*, vol. 49, no. 2, pp. 193–211, Oct. 2017.

[140] K. M. Mahmoud, N. M. Ghanem, and M. A. Ismail, "Unsupervised video summarization via dynamic modeling-based hierarchical clustering," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, vol. 2, 2013, pp. 303–308.

[141] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2069–2077.

[142] G. Guan, Z. Wang, S. Mei, M. Ott, M. He, and D. D. Feng, "A top-down approach for video summarization," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 1, pp. 1–21, Aug. 2014.

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

[144] M. Demir and H. I. Bozma, "Video summarization via segments summary graphs," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1071–1077.

[145] J. Meng, S. Wang, H. Wang, J. Yuan, and Y.-P. Tan, "Video summarization via multiview representative selection," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2134–2145, May 2018.

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

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

[148] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Performance over random: A robust evaluation protocol for video summarization methods," in *Proc. 28th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2020, pp. 1056–1064.

[149] H. S. Chang, S. Sull, and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1269–1279, Dec. 1999.

[150] T. Liu, X. Zhang, J. Feng, and K.-T. Lo, "Shot reconstruction degree: A novel criterion for key frame selection," *Pattern Recognit. Lett.*, vol. 25, no. 12, pp. 1451–1457, 2004.

[151] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[152] S. Kokoska and D. Zwillinger, *CRC Standard Probability and Statistics Tables and Formulae*. Boca Raton, FL, USA: CRC Press, 2000.

[153] C. Collyda, K. Apostolidis, E. Apostolidis, E. Adamantidou, A. I. Metsai, and V. Mezaris, "A web service for video summarization," in *Proc. ACM Int. Conf. Interact. Media Exper.*, New York, NY, USA, Jun. 2020, pp. 148–153.

[154] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1861–1870.

[155] A. Sharghi, B. Gong, and M. Shah, "Query-focused extractive video summarization," in *Proc. ECCV*, 2016, pp. 4788–4797.

[156] A. B. Vasudevan, M. Gygli, A. Volokitin, and L. Van Gool, "Query-adaptive video summarization via quality-aware relevance estimation," in *Proc. 25th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2017, pp. 582–590.

[157] Y. Zhang, M. C. Kampffmeyer, X. Liang, M. Tan, and E. Xing, "Query-conditioned three-player adversarial network for video summarization," in *Proc. Brit. Mach. Vis. Conf.*, Newcastle, U.K.: Northumbria Univ., Sep. 2018, p. 288.

[158] Y. Zhang, M. Kampffmeyer, X. Zhao, and M. Tan, "Deep reinforcement learning for query-conditioned video summarization," *Appl. Sci.*, vol. 9, no. 4, p. 750, Feb. 2019.

[159] J.-H. Huang and M. Worring, "Query-controllable video summarization," in *Proc. Int. Conf. Multimedia Retr.*, New York, NY, USA, Jun. 2020, pp. 242–250, doi: 10.1145/3372278.3390695.

[160] A. G. del Molino, X. Boix, J. Lim, and A. Tan, "Active video summarization: Customized summaries via on-line interaction with the user," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2017, pp. 4046–4052.

[161] A. Ortega, P. Frossard, J. Kovačevič, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[162] Y. Tanaka, Y. C. Eldar, A. Ortega, and G. Cheung, "Sampling signals on graphs: From theory to applications," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 14–30, Nov. 2020.

[163] G. Cheung, E. Magli, Y. Tanaka, and M. K. Ng, "Graph spectral image processing," *Proc. IEEE*, vol. 106, no. 5, pp. 907–930, May 2018.

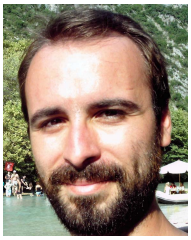
[164] J. H. Giraldozuluaga, S. Javed, and T. Bouwmans, "Graph moving object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, vol. 1, p. 1, Dec. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9288631

[165] T. Doan *et al.*, "On-line adaptative curriculum learning for GANs," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2019, pp. 3470–3477.

[166] K. Ghasedi, X. Wang, C. Deng, and H. Huang, "Balanced self-paced learning for generative adversarial clustering network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4386–4395.

[167] P. Soviany, C. Ardei, R. T. Ionescu, and M. Leordeanu, "Image difficulty curriculum for generative adversarial networks (CuGAN)," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3452–3461.

ABOUT THE AUTHORS



Evlampios Apostolidis received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2007, and the M.Sc. degree in information systems from the University of Macedonia, Thessaloniki, in 2011. He is currently working toward the Ph.D. degree at the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K. His diploma thesis was on methods for digital watermarking of 3-D TV content. For his dissertation, he studied techniques for indexing multidimensional data.

Since January 2012, he has been a Research Assistant with the Centre for Research and Technology Hellas/Information Technologies Institute, Thessaloniki. Since 2018, he has been a Ph.D. student at the School of Electronic Engineering and Computer Science, Queen Mary University of London. He has coauthored three journal articles, four book chapters, and more than 25 conference papers. His research interests lie in the areas of video analysis and understanding, with a particular focus on methods for video segmentation and summarization.

Eleni Adamantidou received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2019, where she graduated in the top 10% of her class (grade 9.01/10).

Since March 2019, she has been a Research Assistant with the Centre for Research and Technology Hellas/Information Technologies Institute, Thessaloniki. She has coauthored one journal article and five conference papers in the field of video summarization. She is particularly interested in deep learning methods for video analysis and summarization, and natural language processing.

Alexandros I. Metsai received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2017. For the needs of his diploma thesis, he developed a robotic agent capable of learning objects shown by a human through voice commands and hand gestures.

Since September 2018, he has been a Research Assistant with the Centre for Research and Technology Hellas/Information Technologies Institute, Thessaloniki. He has coauthored one journal article and four conference papers in the field of video summarization and one book chapter in the field of video forensics. His research interests are in the areas of deep learning for video analysis and summarization.

Vasileios Mezaris (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001 and 2005, respectively.

He is currently a Research Director with the Centre for Research and Technology Hellas/Information Technologies Institute, Thessaloniki. He has coauthored over 40 journal articles, 20 book chapters, 170 conference papers, and three patents. His research interests include multimedia understanding and artificial intelligence, in particular, image and video analysis and annotation, machine learning and deep learning for multimedia understanding and big data analytics, multimedia indexing and retrieval, and applications of multimedia understanding and artificial intelligence.

Dr. Mezaris serves as a Senior Area Editor for IEEE SIGNAL PROCESSING LETTERS and an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA.

Ioannis (Yiannis) Patras (Senior Member, IEEE) is currently a Professor of computer vision and human sensing with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K. He has more than 200 publications in the most selective journals and conferences in the field of computer vision. His research interests lie in the areas of computer vision and human sensing using machine learning methodologies; this includes analysis of human actions and activity in multimedia, affect analysis, and multimodal analysis of human behavior.

Dr. Patras is a member of the Visual Signal Processing and Communications Technical Committee (VSPC) of the IEEE Circuits and Systems (CAS) Society. He is also an Associate Editor of the following journals: *Pattern Recognition and Computer Vision* and *Image Understanding*, and is or has been the Area Chair in all major conferences in the area of computer vision.