Proc. ACM ICMR 2022. Authors' accepted version. The final paper is avaiable at https://doi.org/10.1145/3512527.3531404

Summarizing Videos using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames

Evlampios Apostolidis
CERTH-ITI & Queen Mary University of London
Thessaloniki, Greece, 57001
apostolid@iti.gr

Georgios Balaouras
CERTH-ITI
Thessaloniki, Greece, 57001
mpalaourg@iti.gr

Vasileios Mezaris
CERTH-ITI
Thessaloniki, Greece, 57001
bmezaris@iti.gr

Ioannis Patras
Queen Mary University of London
London, UK, E14NS
i.patras@qmul.ac.uk

ABSTRACT

In this work, we describe a new method for unsupervised video summarization. To overcome limitations of existing unsupervised video summarization approaches, that relate to the unstable training of Generator-Discriminator architectures, the use of RNNs for modeling long-range frames' dependencies and the ability to parallelize the training process of RNN-based network architectures, the developed method relies solely on the use of a self-attention mechanism to estimate the importance of video frames. Instead of simply modeling the frames' dependencies based on global attention, our method integrates a concentrated attention mechanism that is able to focus on non-overlapping blocks in the main diagonal of the attention matrix, and to enrich the existing information by extracting and exploiting knowledge about the uniqueness and diversity of the associated frames of the video. In this way, our method makes better estimates about the significance of different parts of the video, and drastically reduces the number of learnable parameters. Experimental evaluations using two benchmarking datasets (SumMe and TVSum) show the competitiveness of the proposed method against other state-of-the-art unsupervised summarization approaches, and demonstrate its ability to produce video summaries that are very close to the human preferences. An ablation study that focuses on the introduced components, namely the use of concentrated attention in combination with attention-based estimates about the frames' uniqueness and diversity, shows their relative contributions to the overall summarization performance.

CCS CONCEPTS

• **Computing methodologies** → **Video summarization**; **Unsupervised learning**.

ICMR '22, June 27–30, 2022, Newark, NJ, USA

ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531404

KEYWORDS

Video summarization, Unsupervised learning, Concentrated attention, Frame uniqueness, Frame diversity

ACM Reference Format:
Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. 2022. Summarizing Videos using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512527.3531404

1 INTRODUCTION

Video summarization aims to generate a complete and concise synopsis by selecting the most important and informative parts of the full-length video. Current practice in the media industry for the production of a video summary requires an editor to watch the entire content and decide about the parts of it that should be included in the summary. This time-consuming and cumbersome task can be significantly accelerated by technologies for automated video summarization. Hence, such technologies are nowadays of high demand by media organizations, as they can drastically reduce the needed resources for video summary production in terms of both time and human effort.

Several approaches have been proposed to automate video summarization over the last couple of decades, more recently focusing on deep-learning-based methods. This is driven by the successful application of deep network architectures on other video analysis tasks, such as image classification. Based on the reported summarization performance in the relevant works, deep-learning-based video summarization methods represent the current state of the art in the field. A recent study of the literature on deep-learning-based video summarization [4] showed that most works try to learn the summarization task in a supervised manner with the help of ground-truth annotations, but there is also a noticeable number of unsupervised approaches that can be trained without the use of ground-truth data; thus eliminating the need for laborious and time-consuming data annotation tasks. The competitive performance of some of these unsupervised approaches, combined with observations in [4] about the limited amount of available ground-truth data and the heavily-demanding procedures that are required for obtaining such data, highlight the potential of unsupervised video summarization methods.

Proc. ACM ICMR 2022. Authors' accepted version. The final paper is avaiable at https://doi.org/10.1145/3512527.3531404

Driven by the remarks discussed above, we worked towards the development of an unsupervised video summarization method. A study of the relevant literature showed that most existing approaches (e.g., [2, 3, 5, 11–13, 19, 23, 29]) utilize Generator - Discriminator architectures and try to learn how to build a representative summary based on the intuition that such a summary would allow a good reconstruction of the original video. Nevertheless, these methods can suffer from the unstable training of Generative Adversarial Networks (GANs) [34], while their building blocks in most cases are different types of RNNs (mainly LSTMs), which show limited memorization capacity [17]. As an alternative, some works (e.g., [8, 21, 30, 33, 34]) design tailored reward functions that target specific desired characteristics of a good video summary (such as including representative and diverse content), and train a network architecture via reinforcement learning. However, the performance of most of these methods is poor, and they also carry the limitations of using RNNs to model long-range frames' dependencies [17].

To overcome the above discussed drawbacks of existing unsupervised video summarization approaches, we propose a novel approach, called CA-SUM, that relies on the use of a self-attention mechanism. This mechanism involves only matrix multiplication operations that are highly parallelizable, takes into account the entire frame sequence, and can be easily trained in a single forward and backward pass. Moreover, the developed self-attention mechanism is able to concentrate on specific parts of the attention matrix and make better estimates about their significance (this term is used in the sequel as an alternative to the term "attention") by incorporating knowledge about the uniqueness and diversity of the relevant frames of the video. In our method, uniqueness and diversity are somewhat similar notions, where uniqueness is measured by an entropy-based evaluation that involves attention estimates for the entire frame sequence, whereas diversity is computed by assessing the cosine similarity among selected pairs of frames. Finally, our concentrated attention mechanism reduces the number of learnable parameters, and allows the network architecture to be trained efficiently using a simple loss function.

The proposed CA-SUM method follows a completely different approach for learning video summarization, compared to the existing unsupervised methods. It is most closely related to the unsupervised variant of a recently published supervised video summarization algorithm [17], which also relies on the use of a self-attention mechanism that estimates the frames' diversity. However, differently from [17], our method: i) applies a completely different approach for measuring the frames' diversity, ii) extracts and utilizes information also about the frames' uniqueness, iii) and can be trained using a simple loss function that relates to the length of the generated video summary. Our contributions are the following:

- We introduce the use of a concentrated attention mechanism for unsupervised video summarization.
- We propose an approach for evaluating the frames' uniqueness and diversity based on the computed attention values.
- We suggest a method for exploiting the extracted information about the frames' uniqueness and diversity, and produce a block diagonal sparse attention matrix that contains better estimates about the significance of different parts of the video, and reduces the number of learnable parameters.

2 RELATED WORK

Several attempts have been made to automate video summarization. The current state of the art is represented by methods relying on the learning capacity of deep network architectures. For the sake of space, in this section we present in brief the relevant literature on unsupervised video summarization, as the relevant approaches are most closely related to the proposed method. For a more comprehensive survey of the bibliography on deep-learning-based video summarization, interested readers are referred to [4].

The complete absence of any guidance (in the form of ground-truth data) for learning video summarization, led researchers in seeking the most important characteristics of a good video summary. To this direction, most existing unsupervised approaches aim to learn how to build summaries that are highly representative of the video content. Based on the intuition that a representative summary ought to assist the viewer to infer the original video content, most methods rely on the use of Generator-Discriminator architectures along with adversarial learning mechanisms that force the summarization component (which is usually a part of the Generator) to build a summary that allows a good reconstruction of the original video. In a first attempt in this direction, Mahasseni et al. [19] combined an LSTM-based key-frame selector with a Variational Auto-Encoder (VAE) and a trainable Discriminator, and learned video summarization via an adversarial learning process that tries to minimize the distance between the original video and the summary-based reconstructed version of it. Building on the network architecture of [19], Apostolidis et al. proposed a stepwise, label-based approach for training the adversarial part of the network, that leads to improved summarization performance [5]. Further advancement of this performance was achieved in subsequent works of Apostolidis et al. [2, 3]. In [2] the Variational Auto-Encoder was replaced by a deterministic Attention Auto-Encoder in order to learn an attention-driven reconstruction of the original video. In [3] an Actor-Critic (AC) model was embedded in the network architecture of [5] and trained based on a workflow that uses the Discriminator's feedback as a reward and allows the AC model to discover a space of actions and automatically learn a value function (Critic) and a policy for key-fragment selection (Actor). Jung et al. also built on the network architecture of [19] and proposed an extension of it by introducing a chunk and stride network (CSNet) and a tailored difference attention mechanism for assessing the frames' dependence at different temporal granularities [12]. In their following work, Jung et al. [13] replaced the CSNet-based mechanism of [12] for estimating frames' importance, by a self-attention mechanism that is combined with an approach for modeling the relative position between frames. On a similar basis, He et al. [11] integrated a self-attention mechanism in both parts of a Generator-Discriminator architecture. To enhance the frames' importance estimation process, this method uses a feature selector that forces the summarization component of the network to focus on the most important segments of the frame sequence, and models long-range frames' dependencies using multi-head self-attention mechanisms. In another recent work, Wu et al. [29] suggested that taking into account only the representativeness of the video summary is not sufficient for learning the summarization task, as a good summary should also contain the high priority events/entities

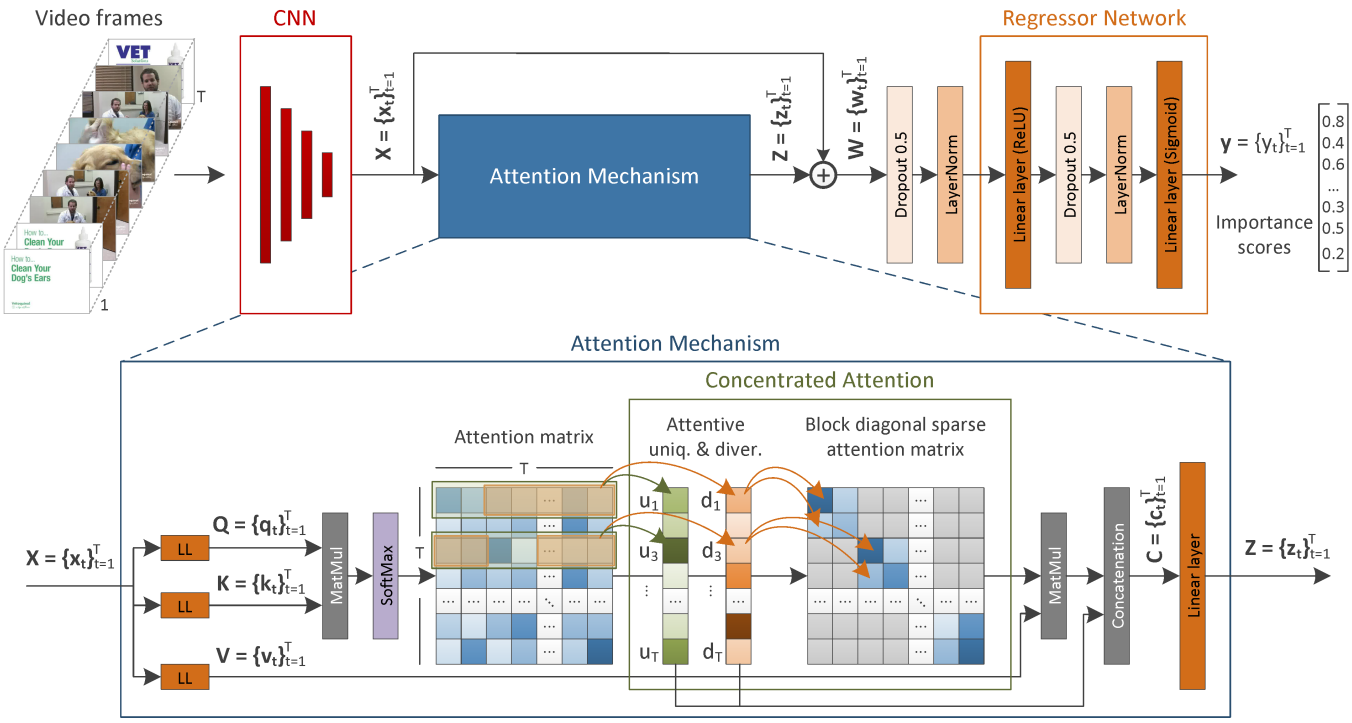

Figure 1: The analysis pipeline of the proposed CA-SUM method. The lower part illustrates the processing steps within the attention mechanism.

of the visual content. Based on this observation, Wu et al. [29] extended a Generator-Discriminator architecture by introducing an Adversarial Spatio-Temporal network that constructs the relationship among entities, using this information while estimating frames' importance, and improving the stability of the Discriminator's training using the earth moving distance of the Wasserstein GAN [6]. On a slightly different basis, Rochan et al. [23] used a Generator-Discriminator architecture to learn video summarization from unpaired data. The generative part is composed of a Fully-Convolutional Sequence Network (FCSN) encoder-decoder, while the discriminative part contains only the decoding part of the FCSN. The goal is 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. Finally, Kanafani et al. [14] focused on the generative part of the network architecture from [12] and investigated the impact of using multiple representations of the visual content while estimating the frames' importance. Different approaches were examined for fusing the computed representations of the visual content at different steps of the analysis and for different temporal granularities.

Taking into account additional desired characteristics for a video summary (besides representativeness, that is targeted by the methods reported in the previous paragraph), a few unsupervised approaches define different hand-crafted reward functions that quantify the existence of desired characteristics in the generated summary, and train a network architecture for video summarization based on reinforcement learning. In this context, Zhou et al. [34] formulated video summarization as a sequential decision-making process and trained a simple LSTM-based network architecture using a pair of diversity and representativeness rewards. The former computes the dissimilarity among the selected key-frames and the latter measures the distance (i.e., the visual resemblance) of the selected key-frames from the remaining video frames. Gonuguntla et al. [8] utilized Temporal Segment Networks [28] to extract spatial and temporal information about the video frames, and trained a video summarization architecture based on a reward function that evaluates the preservation of the video's main spatiotemporal patterns in the produced summary. Zhao et al. [33] described a mechanism that performs both video summarization and reconstruction. Video reconstruction aims to assess the extent to which the summary allows the viewer to infer the original video, and video summarization is learned based on the output of the video reconstruction process and the output of trained models that evaluate the representativeness and diversity of the generated summary. In another recent work, Yaliniz et al. [30] used Independently Recurrent Neural Networks (IndRNNs) [18] to model the temporal dependence of video frames, and learned summarization by using rewards associated with the representativeness, diversity and uniformity (i.e., the temporal coherence) of the video summary. Finally, Phaphuangwittayakul et al. [21] presented a variation of the network architecture from [34], which estimates the frames' importance by combining the created representations for the video frames at the output of a bi-directional RNN and a self-attention mechanism.

Besides the unsupervised video summarization approaches discussed above, a few variations of supervised methods that can be trained without using ground-truth annotations have appeared in the literature [17, 24, 32]. These variations are also taken into account in the performance comparisons reported in Section 4.

3 PROPOSED APPROACH

The basis of our developments was the network architecture of VASNet [7]. The heart of this supervised video summarization method is a soft self-attention mechanism that takes into account the entire frame sequence and models the frames' dependencies according to their pair-wise similarities in a learned latent space. The output of this mechanism is forwarded to a two-layer, fully connected network that produces estimates about each frame's importance. These estimates are compared with ground-truth annotations about the frames' importance, and the computed loss value guides the supervised training of the architecture.

To train this architecture in a fully-unsupervised manner, we developed a new method that allows the attention mechanism to: i) concentrate on specific parts of the attention matrix, that correspond to different non-overlapping video fragments of fixed length, ii) make better estimates about the significance of each of these parts, by extracting and exploiting information about the uniqueness and diversity of the associated video frames, iii) produce a block diagonal sparse attention matrix, thus significantly reducing the number of learnable parameters, and iv) learn the video summarization task via a simple loss function that relates to the length of the generated summary. In the following, we present the processing pipeline of the proposed method (Fig. 1), from video representation to frames' importance estimation. This pipeline is the same during both training and inference. Regarding the used notation: capital bold letters denote matrices, small bold letters denote vectors and non-bold letters (either capital or small) denote scalars.

Given a video of T frames, the CA-SUM method initially produces a set of deep feature representations ($\boldsymbol{X} = \{\boldsymbol{x_t}\}_{t=1}^{T}$) of size D ($\boldsymbol{x_t} = \{x_{t,r}\}_{r=1}^{D}$) using a pretrained CNN model. These representations form the input of the attention mechanism and are utilized also via a residual skip connection to facilitate back-propagation and assist the model's convergence. As illustrated in the lower part of Fig. 1, the attention mechanism passes the set of frame feature vectors ($\boldsymbol{X} = \{\boldsymbol{x_t}\}_{t=1}^{T}$) through three different linear layers and forms the Query ($\boldsymbol{Q} = \{\boldsymbol{q_t}\}_{t=1}^{T}$), Key ($\boldsymbol{K} = \{\boldsymbol{k_t}\}_{t=1}^{T}$) and Value ($\boldsymbol{V} = \{\boldsymbol{v_t}\}_{t=1}^{T}$) matrices of the training process. The Query and Key matrices participate in a matrix multiplication process that is followed by a softmax layer, and formulate the values of the attention matrix ($\boldsymbol{A} = \{a_{i,j}\}_{i,j=1}^{T}$). The computed attention values are then utilized by the concentrated attention mechanism. This mechanism focuses on non-overlapping blocks of size M that are in the main diagonal of the attention matrix, and aims to enrich the existing information in each of these blocks by integrating knowledge about the uniqueness and diversity of the associated frames of the video. The different processing steps of the concentrated attention mechanism, are presented in Alg. 1. The outputs of this mechanism are: i) a block diagonal sparse attention matrix that concentrates the information about the significance of different consecutive and non-overlapping parts of the video in a narrow area around its main diagonal, and ii) a pair of values for each frame, that represent its attentive uniqueness ($u_t \in \mathbb{R}^+$ and $t \in [1, T]$) and diversity ($d_t \in \mathbb{R}^+$ and $t \in [1, T]$) from the video frames outside the block of interest. The adopted terminology for these two values relates to the fact that the computed attention values ($\boldsymbol{A} = \{a_{i,j}\}_{i,j=1}^{T}$) are taken into account for estimating the frames' uniqueness and diversity. The formulas for calculating these values are described in Alg. 1. Each pair of attentive uniqueness and diversity values is concatenated at the end of the corresponding feature vector that is created after a matrix multiplication process involving the block diagonal sparse attention matrix ($\boldsymbol{B} = \{b_{i,j}\}_{i,j=1}^{T}$) and the Value matrix ($\boldsymbol{V} = \{\boldsymbol{v_t}\}_{t=1}^{T}$). The generated feature vectors ($\boldsymbol{C} = \{\boldsymbol{c_t}\}_{t=1}^{T}$) of size $D + 2$ ($\boldsymbol{c_t} = \{c_{t,r}\}_{r=1}^{D+2}$) pass through a linear layer that reduces their dimensionality to D. The generated feature vectors at the

Algorithm 1 The processing steps of the concentrated attention mechanism.

Notation: T is the number of video frames, N is the number of blocks, M is the block size, $\boldsymbol{A}$ is the attention matrix, $\boldsymbol{B}$ is the block diagonal sparse attention matrix, u_t and d_t are the attentive uniqueness and diversity values for the t^{th} frame respectively.

Input: The attention matrix $\boldsymbol{A}$.

Output: The block diagonal sparse attention matrix $\boldsymbol{B}$, and the computed values about the attentive uniqueness (u_t) and diversity (d_t) for each video frame ($t \in [1, T]$).

"""

Estimate the attentive uniqueness of each video frame by computing the entropy of each row of the attention matrix $\boldsymbol{A}$*, using the following formula:* $Entropy(\boldsymbol{a_i}) = -\sum_{t=1}^{T} a_{i,t} \cdot log(a_{i,t})$

"""

for $i = 1 \rightarrow T$ **do**
  $e_i = Entropy(\boldsymbol{a_i})$
$\boldsymbol{u} = ||\boldsymbol{e}||_1$

"""

For each block of the block diagonal sparse attention matrix, estimate the attentive diversity of each frame of the block by computing the mean of its attention-based weighted dissimilarities (Dis) from the frames that lie outside the block.

"""

for $k = 0 \rightarrow N - 1$ **do**
  for $i = Mk + 1 \rightarrow M(k + 1)$ **do**
    $IND = [Mk + 1, M(k + 1)]$ *# indices inside block*
    for $l = 1 \rightarrow T$ & $(l \notin IND)$ **do**
      $Dis(i, l) = 1 - \frac{\boldsymbol{x_i} \cdot \boldsymbol{x_l^T}}{||\boldsymbol{x_i}||_2 \cdot ||\boldsymbol{x_l}||_2}$ *# cosine distance*
    $d_i = \frac{1}{\sum_l Dis(i,l)} \sum_l (Dis(i, l) \cdot a_{i,l})$

"""

Compute the block diagonal sparse attention matrix $\boldsymbol{B}$

"""

$\boldsymbol{B} = zeros(T, T)$ *# create a TxT zero matrix*
for $k = 0 \rightarrow N - 1$ **do**
  for $j = Mk + 1 \rightarrow M(k + 1)$ **do**
    for $i = Mk + 1 \rightarrow M(k + 1)$ **do**
      $b_{i,j} = a_{i,j} + d_j$

output of the attention mechanism ($\boldsymbol{Z} = \{\boldsymbol{z_t}\}_{t=1}^{T}$) are then added to the original feature vectors ($\boldsymbol{X} = \{\boldsymbol{x_t}\}_{t=1}^{T}$) via a residual skip connection (this addition is represented by the $\oplus$ symbol in the upper part of Fig. 1). The result of this operation ($\boldsymbol{W} = \{\boldsymbol{w_t}\}_{t=1}^{T}$) is forwarded to a dropout layer that is followed by a normalization layer. The produced representations are given as input to the Regressor Network, which is the same with the one in [7]. Finally, the Regressor produces a set of frame-level scores ($\boldsymbol{y} = \{y_t\}_{t=1}^{T}$ with $y_t \in \mathbb{R}$ and $y_t \in (0, 1)$) that indicate the frames' importance.

Given the output of the aforementioned processing pipeline, at training time, we compute a length regularization loss:

$$L_{reg} = \left| \frac{1}{T} \sum_{t=1}^{T} y_t - \sigma \right|, \tag{1}$$

where σ is the summary length regularization factor, a tunable hyper-parameter of our method (σ was introduced in [19] and is used in several subsequent works, e.g., [3, 17, 21, 34]). The computed training loss is then back-propagated to compute the gradients and update all the different parts of the architecture.

At inference time, the estimated importance scores are used to select the key-fragments of the video and form the video summary. For this, given a temporal segmentation of the video into its building blocks (obtained e.g., using the KTS algorithm [22]), fragment-level importance is calculated by averaging the scores of the frames within each fragment. Finally, requiring that the summary does not exceed 15% of the video duration (which is a common evaluation-protocol setting in the relevant literature), we form the video summary by solving the Knapsack problem, similarly to [2, 3, 5, 8, 11–14, 17, 19, 21, 23, 29, 30, 32–34].

4 EXPERIMENTS

4.1 Datasets and evaluation approach

Datasets. The performance of the proposed method is evaluated on two benchmarking datasets. The SumMe dataset [10] includes 25 videos (1-6 min. duration) with diverse video contents (e.g., covering holidays, events and sports), captured from both first-person and third-person view. Each video has been annotated by 15 to 18 users in the form of key-fragments, and thus is associated to multiple user summaries of varying length (5-15% of the initial video duration). The TVSum dataset [26] contains 50 videos (1-11 min. duration) from 10 categories of the TRECVid MED task [25] (5 videos per category). Each video has been annotated by 20 users in the form of frame-level importance scores, ranging from 1 (not important) to 5 (very important).

Evaluation Approach. Our assessments are based on two different evaluation approaches. The first approach (proposed in [31] and adopted by the majority of the state of the art video summarization methods) estimates the similarity between a machine-generated (M) and a user-defined summary (U) by computing their overlap using the F-Score (as percentage), where (P)recision and (R)ecall measure the temporal overlap ($\cap$) between the summaries at the frame-level ($||*||$ denotes duration):

$$F = 2 \times \frac{P \times R}{P + R} \times 100, \quad \text{with} \quad P = \frac{M \cap U}{||M||} \quad \text{and} \quad R = \frac{M \cap U}{||U||} \tag{2}$$

This computation is directly feasible for the videos of the SumMe dataset, as their annotations are in the form of key-fragments. For the videos of the TVSum dataset, the available frame-level annotations are initially converted to key-fragment annotations by applying the methodology described in [26, 31]. So, for a given test video we compare the generated summary with the available user summaries for this video, and compute an F-Score for each pair of compared summaries. Then, we average the computed F-Scores (for the TVSum videos, as proposed in [26]) or keep the maximum of them (for the SumMe videos, as suggested in [9] to account for the fact that multiple summaries of varying length are possible for a video) and form the final F-Score for this video. The computed F-Scores for all test videos are averaged and this average indicates the method's performance on the test set.

The second evaluation approach (proposed in [20]) eliminates the impact of any utilized video fragmentation and key-fragment selection mechanism (e.g., the Knapsack algorithm). It assesses the quality of a machine-generated video summary by considering the produced frame-level importance scores at the output of a network architecture as rankings, and comparing them with the user-generated frame-level importance scores using the Kendall's τ [15] and Spearman's ρ [16] rank correlation coefficients. In this case, for a given test video we compare the estimated frame-level importance scores with the available user annotations (also frame-level importance scores) for this video, and compute the τ and ρ values for each pair of comparisons. Then, we average the computed sets of τ and ρ values, and form the final τ and ρ values for this video. The computed τ and ρ values for all test videos are then averaged and this average defines the method's performance on the test set. This evaluation approach is applicable only on the videos of the TVSum dataset, that have been annotated with frame-level importance scores.

Finally, let us note that similarly to most of the existing works (e.g., [2, 3, 5, 7, 8, 19, 21, 23, 29, 32, 34]) we split each dataset into a training (80% of the videos) and a testing set (the remaining 20% of the videos) and we run experiments on five different randomly-created splits for each dataset. In the following, we report the average performance over these runs (cf. Tables 1-3 and 5).

4.2 Implementation details

For fair comparisons with the literature, videos are downsampled to 2 fps, and deep representations of size $D = 1024$ are obtained for the sampled frames by taking the output of the pool5 layer of GoogleNet [27] trained on ImageNet. The block size M is set to 60, based on the findings of a sensitivity analysis that examined different options for the value of parameter M (see Section 4.3). The learning rate and the L2 regularization factor for training the architecture are equal to $5 \cdot 10^{-4}$, and 10^{-5}, respectively. For initializing the network's parameters we use the Xavier uniform initialization approach with gain = $\sqrt{2}$ and biases = 0.1 (as in [7]). Training is performed in a full-batch mode (i.e., batch size is equal to the number of training samples) using the Adam optimizer, and stops after 400 epochs.

Based on the observations of Mahasseni et al. [19] regarding the impact of the length regularization factor σ on the summarization performance, instead of manually choosing a value for this hyper-parameter we consider several values ranging in $[0.5, 0.9]$

with a step equal to 0.1. After the end of the training, we apply a model selection criterion that is responsible for selecting a well-trained model by indicating both the training epoch and the value of the length regularization factor σ. In particular, we initially keep one trained model per considered σ value, by selecting the epoch that corresponds to the minimum training loss. Then, our decision on which one of these five models performs the best is based on transductive inference. In particular, we extend the set of selected models by adding an untrained model of our CA-SUM network architecture (i.e., a model with random weights), and we apply the entire set of the six models on the videos of the test set. Then, we select the model that shows the best improvement compared to the untrained model, normalized by the progress that should be made by the untrained model for producing frame importance scores that average towards the value of the length regularization factor σ. More specifically, for each one of the trained models, we compute the following score:

$$S = \left| \frac{\mu_{tr} - \mu_{un}}{\sigma - \mu_{un}} \right|, \tag{3}$$

where μ_{tr} and μ_{un} correspond to the mean value of the produced importance scores for the video frames over the entire set of test videos, when using a trained and the untrained model respectively, and are defined as follows:

$$\mu_{un} = \frac{1}{F} \sum_{j=1}^{F} \frac{1}{T_j} \sum_{i=1}^{T_j} y_i^{un} \quad \text{and} \quad \mu_{tr} = \frac{1}{F} \sum_{j=1}^{F} \frac{1}{T_j} \sum_{i=1}^{T_j} y_i^{tr}, \tag{4}$$

where F denotes the number of test videos, T_j indicates the number of frames of the j^{th} test video, and $\boldsymbol{y^{un}}$, $\boldsymbol{y^{tr}}$ denote the computed importance scores for the frames of the current test video by the untrained and the trained model, respectively. Given this set of scores, we select as the best-performing model (i.e., the value of the hyper-parameter σ) the one with an S score that is the closest to an upper-bounding experimentally-defined threshold $\xi = 1.5$, where values of S greater than this threshold indicate overfitting.

All experiments were carried out on a PC with an NVIDIA RTX 2080Ti GPU. To allow the reproduction of our results, the PyTorch implementation of CA-SUM is publicly-available at: https://github.com/e-apostolidis/CA-SUM.

4.3 Sensitivity analysis

Since the concentrated attention mechanism, which produces the block diagonal sparse attention matrix, is in the core of the processing pipeline at both training and inference stage, we examined different options about the size M of the block. This size indicates the length of the video fragment where the attention mechanism concentrates each time, and thus the level of granularity for estimating the attention-based significance of the video. The different options for this hyper-parameter, and the summarization performance of the CA-SUM model according to the F-Score, Spearman's ρ and Kendall's τ measures, are presented in Table 1. These results show that the increment of the block size almost constantly leads to improved summarization performance on TVSum, but the findings for the SumMe dataset are mixed. The best option for this hyper-parameter is $M = 60$ as it leads to the highest performance on both datasets and according to all considered measures. Higher values were not taken under consideration in this experiment, as the maximum possible value is dictated by the shortest video in our datasets, which in our case is 60 frames. Finally, the use of blocks that correspond to the detected video fragments by the KTS algorithm (which is the most commonly used approach for video segmentation in the relevant literature [2, 3, 5, 8, 11–14, 17, 19, 21, 24, 29, 30, 32–34]) results in the worst summarization performance.

Table 1: The performance of the CA-SUM method on the SumMe and TVSum dataset, for different options about the block size and according to the overlap between the machine and the human-generated summaries (F-Score (%)), and based on the comparison of the machine and the human-defined importance scores for the video frames (Spearman's ρ and Kendall's τ rank correlation coefficients).

Block size	SumMe	TVSum	TVSum	
	F-Score		Spearman's ρ	Kendall's τ
10	44.8	55.7	0.090	0.070
20	46.3	57.0	0.140	0.110
30	47.0	57.6	0.160	0.120
40	46.1	60.5	0.170	0.130
50	44.3	60.4	0.190	0.150
60	**51.1**	**61.4**	**0.210**	**0.160**
KTS	40.8	52.7	0.060	0.050

4.4 Performance comparisons

The proposed CA-SUM method was compared with a random summarizer and several state-of-the-art unsupervised video summarization approaches. The performance of a random summarizer on a given video was measured as proposed in [1]. In particular, we initially assigned randomly-created importance scores to the video frames based on a uniform distribution of probabilities. Then, we computed fragment-level scores based on a predefined KTS-based segmentation of the video. Finally, we used the Knapsack algorithm to form a summary with a length that does not exceed 15% of video duration. Random summarization was performed 100 times and we report the average score over these runs. The performance of each compared unsupervised method is from the corresponding paper, unless stated otherwise. The reported values in Table 2 show that the proposed CA-SUM method performs consistently good on both of the utilized datasets, being the top-performing one on TVSum and the second best performing one on SumMe, according to the F-Score measure. In addition, our idea to work with a self-attention mechanism seems to be right, as five out of the six top-performing approaches ([12, 13, 17, 21]) contain some kind of attention mechanism. The use of more simple versions of Generator-Discriminator architectures ([2, 5, 29]) cannot lead to competitive performance on the used datasets, given the current state of the art. Finally, the integration of a self-attention mechanism into a network architecture that is trained based on reinforcement learning ([21]) leads to significantly higher performance compared to simpler architectures that adopt the same learning approach ([8, 34]).

The improved summarization performance of the proposed CA-SUM method is highlighted also by the results shown in Table 3.

Table 2: Comparison with unsupervised video summarization approaches on SumMe and TVSum. F1 denotes F-Score (%) and Rnk denotes the ranking of the compared methods. Approaches marked with a star (⋆) have been evaluated using the same five splits of data.

	SumMe		TVSum		Avg	Data
	F1	Rnk	F1	Rnk	Rnk	splits
Random summary	40.2	18	54.4	15	16.5	–
SUM-FCN$_{unsup}$ [24]	41.5	16	52.7	16	16	M Rand
DR-DSN [34]	41.4	17	57.6	12	14.5	5 Rand [1]
EDSN [8]	42.6	14	57.3	13	13.5	5 Rand [2]
RSGN$_{unsup}$ [32]	42.3	15	58.0	11	13	5 Rand
UnpairedVSN [23]	47.5	11	55.6	14	12.5	5 Rand
PCDL [33]	42.7	13	58.4	9	11	5 FCV
ACGAN [11]	46.0	12	58.5	8	10	5 FCV
⋆SUM-Ind_{LU} [30]	46.0	12	58.7	7	9.5	5 Rand [3]
⋆ERA [29]	48.8	8	58.0	11	9.5	5 Rand
⋆SUM-GAN-sl [5]	47.8	10	58.4	9	9.5	5 Rand
⋆SUM-GAN-AAE [2]	48.9	7	58.3	10	8.5	5 Rand
⋆MCSF$_{late}$ [14]	47.9	9	59.1	5	7	5 Rand
SUM-GDA$_{unsup}$ [17]	50.0	6	59.6	4	5	5 FCV
CSNet+GL+RPE [13]	50.2	5	59.1	5	5	5 FCV
CSNet [12]	**51.3**	1	58.8	6	3.5	5 FCV
⋆DSR-RL-GRU [21]	50.3	4	60.2	3	3.5	5 Rand
⋆AC-SUM-GAN [3]	50.8	3	60.6	2	2.5	5 Rand
⋆CA-SUM (Proposed)	51.1	2	**61.4**	1	**1.5**	5 Rand

Table 3: Comparison with unsupervised video summarization approaches on TVSum, using the rank correlation coefficients proposed in [20].

	Spearman's ρ	Kendall's τ
Random summary [20]	0.000	0.000
Human summary [20]	0.204	0.177
DR-DSN [34]	0.026	0.020
CSNet [12]	0.034	0.025
RSGN$_{unsup}$ [32]	0.052	0.048
CSNet+GL+RPE [13]	0.091	0.070
DSR-RL-GRU [21]	0.114	0.086
CA-SUM (Proposed)	**0.210**	**0.160**

Our method is by far the top performing one, according to both Spearman's ρ and Kendall's τ rank correlation coefficients. Moreover, the performance of CA-SUM on the used data splits of the TVSum dataset approximates the performance of the average human annotator. These results clearly indicate that the proposed method is able to produce structures of frame-level importance scores, that are highly aligned to the human preferences.

[1] The evaluation of this method was made using 5 randomly-created data splits, according to the publicly-released code by the authors of this work.

[2] The evaluation of this method was made according to the protocol adopted in [34].

[3] In the original work this method was evaluated based on a variation of the established evaluation approach, which compares the machine-generated summary with the single ground-truth summary that is available for each video of the utilized datasets (purely for supervised training). The results reported here are from the work in [29], which assessed the performance of the SUM-Ind_{LU} method according to the established evaluation approach.

Table 4: Comparison of different unsupervised video summarization methods with publicly available implementations, with respect to the training time (seconds per training epoch) and the amount of learnable parameters (in Millions).

Method	Training time (sec / epoch)		# Parameters (in Millions)
	SumMe	TVSum	
DR-DSN [34]	0.33	0.98	2.63
SUM-Ind_{LU} [30] [4]	2.07	9.84	0.33
SUM-GAN-sl [5]	11.85	38.95	23.31
SUM-GAN-AAE [2]	16.39	54.23	24.31
CSNet [12] [4]	28.43	89.85	100.76
DSR-RL-GRU [21]	0.23	0.50	13.64
AC-SUM-GAN [3]	28.25	93.80	26.75
CA-SUM (Proposed)	0.06	0.13	5.25

To investigate how well suited the produced video summaries are for human consumption, we extracted some statistics about the selected video fragments for creating these summaries (note: in the sequel, the values in parentheses represent standard deviation). The applied video summarization pipeline (described in the last paragraph of Section 3) selects on average 27.85%(±3.21) and 32.93%(±4.50) of the video fragments of the SumMe and TVSum videos, respectively. The mean duration of the selected fragments for the SumMe videos is 2.62(±0.69) seconds and for the TVSum videos is 2.26(±0.68) seconds. Given the fact that the average duration of the entire set of video fragments for the videos of the SumMe and TVSum dataset is 4.91(±2.58) and 4.93(±3.40) seconds respectively, we can see that the applied video summarization pipeline selects fragments that are shorter than the average fragment length (approximately half of it). However, we should point out that this selection is significantly affected by the applied Knapsack algorithm, which promotes the selection of short video fragments while trying to maximize the overall importance of the video summary for a given time budget.

Finally, we compared different unsupervised video summarization methods with publicly available implementations, with respect to the training time (seconds per training epoch) and the amount of learnable parameters (in Millions). These experiments were carried out using the same PC (with an i5-11600K CPU, 32GB RAM and an NVIDIA RTX 2080 Ti GPU) and the exact same five splits of data. The findings presented in Table 4 show that the use of Generator-Discriminator architectures for learning the summarization task ([2, 3, 5, 12]) is the most demanding approach, in terms of both the training time and the memory required. The use of less complex architectures where the estimation of the frames' importance is based on the modeling of the frames' dependencies purely using RNNs ([30, 34]), can significantly reduce the training time, but at the cost of a lower summarization performance (as shown in Table 2). Finally, the use of self-attention mechanisms - either solely (as in our proposed method) or in combination with RNNs (as in [21]) - leads to several times faster training, while ensuring high summarization performance. Being highly parallelizable, the proposed

[4] Code re-implemented and publicly-released by the authors of [14].

Table 5: Ablation study based on the performance (F-Score (%), Spearman's ρ, Kendall's τ) of different variants of the proposed model, on SumMe and TVSum.

	Block diagonal sparse attention matrix	Attentive frame uniq. & diver.	SumMe	TVSum	TVSum	
			F-Score		Spearman's ρ	Kendall's τ
Variant #1	X	X	45.8	58.9	NaN	NaN
Variant #2	✓	X	47.4	56.5	0.010	0.010
Variant #3	X	✓	45.8	58.9	NaN	NaN
CA-SUM w/o threshold ξ	✓	✓	49.3	61.2	0.200	0.150
CA-SUM (Proposed)	✓	✓	**51.1**	**61.4**	**0.210**	**0.160**

CA-SUM method needs the least time for model training; and despite the fact that we train five different models (one per different value of the length regularization factor σ), the overall time needed for training still remains at the very low levels, being very close to the time required by the fastest method (DSR-RL-GRU) among the remaining ones. Finally, with respect to the required computational resources for training the proposed CA-SUM method, the memory footprint of the network architecture is 1.16 GBs; this means that our method can process up to 2 hours of video content (in a single-batch mode) using a GPU with memory capacity similar to that of an NVIDIA RTX 2080 Ti. Nevertheless, working with significantly larger videos than the ones included in the utilized datasets, would require the tuning of some hyper-parameters of our method, such as the block size M.

4.5 Ablation study

To assess the impact of each of the main changes that were introduced in the processing pipeline of the supervised VASNet method [7] in order to develop our unsupervised CA-SUM model, we conducted an ablation study that included the following variants of the proposed architecture:

- Variant #1 leaves out the entire concentrated attention mechanism, thus avoiding the computation of both the block diagonal sparse attention matrix and the estimates about the attentive uniqueness and diversity of the video frames. This variant can be considered as an unsupervised version of the VASNet method [7].
- Variant #2 does not compute any estimates about the frames' uniqueness and diversity, and the created block diagonal sparse attention matrix can be seen as the case where the attention mechanism performs local attention only.
- Variant #3 excludes the generation of a block diagonal sparse attention matrix, and the computed estimates about the frames' attentive uniqueness and diversity are used only to enrich the output of a matrix multiplication process that involves the Attention ($\boldsymbol{A}$) and the Value ($\boldsymbol{V}$) matrices.

Besides the above described variants, we also examined the performance of a variation of our CA-SUM method, that does not use any threshold ξ during model selection. The results in Table 5 show that removing the block diagonal sparse attention matrix (Variants #1 and #3) does not allow the network to learn anything meaningful. In particular, the network sticks in a case where all the video frames are assigned with the same extremely low (≈ 0) or extremely high (≈ 1) importance scores. The generation of the summary completely depends on the choices made by the Knapsack algorithm, that are purely related to the length of the video fragments. Since there is no counted variation in the assigned importance scores, no values can be computed for the Spearman's ρ and Kendall's τ rank correlation coefficients. In the case where no estimates about the frames' attentive uniqueness and diversity are computed, and the block diagonal sparse attention matrix provides information about the frames' attention only at the local-level (Variant #2), the network also fails to learn the summarization task. The computed F-Score, Spearman's ρ and Kendall's τ values, indicate a random-level performing summarizer on both utilized datasets. The combination of a block diagonal sparse attention matrix with attention-based statistics about the frames' uniqueness and diversity, as proposed, allows the network architecture to effectively learn the video summarization task and exhibit state-of-the-art performance on both of the utilized datasets and according to both of the adopted evaluation protocols. Finally, the variation of our method that excludes the threshold ξ for model selection, maintains the high levels of video summarization performance on TVSum and performs competitively on SumMe.

5 CONCLUSION

In this paper, we proposed a new method for unsupervised video summarization, that aims to overcome drawbacks of existing approaches with respect to: i) unstable training of Generator - Discriminator architectures, ii) the use of RNNs for modeling long-range frames' dependencies, and iii) the parallelization ability of the training process of existing RNN-based network architectures. The developed CA-SUM method relies on the use of a self-attention mechanism, and extends its functionality by concentrating attention on non-overlapping blocks in the main diagonal of the attention matrix and by utilizing information about the frames' uniqueness and diversity. The proposed methodology allows the attention mechanism to make better estimates about the importance of different parts of the video. Experiments on two benchmarking datasets (SumMe and TVSum) indicated the competitiveness of our CA-SUM method against other state-of-the-art unsupervised summarization approaches, and demonstrated its ability to produce summaries that meet the human expectations.

ACKNOWLEDGMENTS

This work was supported by the EU Horizon 2020 programme under grant agreements H2020-832921 MIRROR and H2020-951911 AI4Media.

REFERENCES

[1] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2020. Performance over Random: A Robust Evaluation Protocol for Video Summarization Methods. In *Proc. of the 28th ACM Int. Conf. on Multimedia (MM '20)* (Seattle, WA, USA). ACM, New York, NY, USA, 1056–1064. https://doi.org/10.1145/3394171.3413632

[2] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2020. Unsupervised Video Summarization via Attention-Driven Adversarial Learning. In *Proc. of the 26th Int. Conf. on Multimedia Modeling (MMM 2020)*. Springer International Publishing, Cham, 492–504. https://doi.org/10.1007/978-3-030-37731-1_40

[3] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 8 (2021), 3278–3292. https://doi.org/10.1109/TCSVT.2020.3037883

[4] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video Summarization Using Deep Neural Networks: A Survey. *Proceedings of the IEEE* 109, 11 (2021), 1838–1863. https://doi.org/10.1109/JPROC.2021.3117472

[5] Evlampios Apostolidis, Alexandros I. Metsai, Eleni Adamantidou, Vasileios Mezaris, and Ioannis Patras. 2019. A Stepwise, Label-based Approach for Improving the Adversarial Training in Unsupervised Video Summarization. In *Proc. of the 1st Int. Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV '19)* (Nice, France). ACM, New York, NY, USA, 17–25. https://doi.org/10.1145/3347449.3357482

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proc. of the 34th Int. Conf. on Machine Learning*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, 214–223. https://proceedings.mlr.press/v70/arjovsky17a.html

[7] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2019. Summarizing Videos with Attention. In *Asian Conf. on Computer Vision (ACCV) 2018 Workshops*, Gustavo Carneiro and Shaodi You (Eds.). Springer International Publishing, Cham, 39–54.

[8] N Gonuguntla, B Mandal, NB Puhan, et al. 2019. Enhanced Deep Video Summarization Network. In *Proc. of the 2019 British Machine Vision Conf. (BMVC)*.

[9] Michael Gygli, Helmut Grabner, and Luc Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 3090–3098. https://doi.org/10.1109/CVPR.2015.7298928

[10] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating Summaries from User Videos. In *Europ. Conf. on Computer Vision (ECCV) 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 505–520.

[11] Xufeng He, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. 2019. Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks. In *Proc. of the 27th ACM Int. Conf. on Multimedia (MM '19)* (Nice, France). ACM, New York, NY, USA, 2296–2304. https://doi.org/10.1145/3343031.3351056

[12] Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. 2019. Discriminative feature learning for unsupervised video summarization. In *Proc. of the 2019 AAAI Conf. on Artificial Intelligence*.

[13] Yunjae Jung, Donghyeon Cho, Sanghyun Woo, and In So Kweon. 2020. Global-and-Local Relative Position Embedding for Unsupervised Video Summarization. In *Europ. Conf. on Computer Vision (ECCV) 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 167–183.

[14] Hussain Kanafani, Junaid Ahmed Ghauri, Sherzod Hakimov, and Ralph Ewerth. 2021. Unsupervised Video Summarization via Multi-Source Features. In *Proc. of the 2021 Int. Conf. on Multimedia Retrieval* (Taipei, Taiwan) *(ICMR '21)*. Association for Computing Machinery, New York, NY, USA, 466–470. https://doi.org/10.1145/3460426.3463597

[15] Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika* 33, 3 (1945), 239–251.

[16] Stephen Kokoska and Daniel Zwillinger. 2000. *CRC standard probability and statistics tables and formulae*. Crc Press.

[17] Ping Li, Qinghao Ye, Luming Zhang, Li Yuan, Xianghua Xu, and Ling Shao. 2021. Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition* 111 (2021), 107677.

[18] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. 2018. Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. In *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 5457–5466. https://doi.org/10.1109/CVPR.2018.00572

[19] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised Video Summarization with Adversarial LSTM Networks. In *2017 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2982–2991.

[20] Mayu Otani, Yuta Nakahima, Esa Rahtu, and Janne Heikkilä. 2019. Rethinking the Evaluation of Video Summaries. In *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[21] Aniwat Phaphuangwittayakul, Yi Guo, Fangli Ying, Wentian Xu, and Zheng Zheng. 2021. Self-Attention Recurrent Summarization Network with Reinforcement Learning for Video Summarization Task. In *Proc. of the 2021 IEEE Int. Conf. on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME51207.2021.9428142

[22] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-Specific Video Summarization. In *Europ. Conf. on Computer Vision (ECCV) 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 540–555.

[23] Mrigank Rochan and Yang Wang. 2019. Video Summarization by Learning From Unpaired Data. In *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 7894–7903.

[24] Mrigank Rochan, Linwei Ye, and Yang Wang. 2018. Video Summarization Using Fully Convolutional Sequence Networks. In *Europ. Conf. on Computer Vision (ECCV) 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 358–374.

[25] Alan F. Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation Campaigns and TRECVid. In *Proc. of the 8th ACM Int. Workshop on Multimedia Information Retrieval* (Santa Barbara, California, USA) *(MIR '06)*. Association for Computing Machinery, New York, NY, USA, 321–330. https://doi.org/10.1145/1178677.1178722

[26] Yale Song, J. Vallmitjana, A. Stent, and A. Jaimes. 2015. TVSum: Summarizing web videos using titles. In *2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 5179–5187. https://doi.org/10.1109/CVPR.2015.7299154

[27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[28] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Europ. Conf. on Computer Vision (ECCV) 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 20–36.

[29] Guande Wu, Jianzhe Lin, and Cláudio T. Silva. 2021. ERA: Entity Relationship Aware Video Summarization with Wasserstein GAN. In *Proc. of the 2021 British Machine Vision Conf. (BMVC)*.

[30] Gokhan Yaliniz and Nazli Ikizler-Cinbis. 2021. Using independently recurrent networks for reinforcement learning based unsupervised video summarization. *Multimedia Tools and Applications* 80, 12 (2021), 17827–17847. https://doi.org/10.1007/s11042-020-10293-x

[31] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video Summarization with Long Short-Term Memory. In *Europ. Conf. on Computer Vision (ECCV) 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 766–782.

[32] Bin Zhao, Haopeng Li, Xiaoqiang Lu, and Xuelong Li. 2021. Reconstructive Sequence-Graph Network for Video Summarization. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3072117

[33] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2020. Property-Constrained Dual Learning for Video Summarization. *IEEE Trans. on Neural Networks and Learning Systems* 31, 10 (2020), 3989–4000. https://doi.org/10.1109/TNNLS.2019.2951680

[34] Kaiyang Zhou and Yu Qiao. 2018. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. In *Proc. of the 2018 AAAI Conf. on Artificial Intelligence*.