EXPLOITING CAPTION DIVERSITY FOR UNSUPERVISED VIDEO SUMMARIZATION

Michail Kaseris *Ioannis Mademlis* *Ioannis Pitas*

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

ABSTRACT

Most unsupervised Deep Neural Networks (DNNs) for video summarization rely on adversarial learning, autoencoding and training without utilizing any ground-truth summary. In several cases, the Convolutional Neural Network (CNN)-derived video frame representations are sequentially fed to a Long Short-Term Memory (LSTM) network, which selects key-frames and, during training, attempts to reconstruct the original/full video from the summary, while confusing an adversarially optimized Discriminator. Additionally, regularizers aiming at maximizing the summary's visual semantic diversity can be employed, such as the Determinantal Point Process (DPP) loss term. In this paper, a novel DPP-based regularizer is proposed that exploits a pretrained DNN-based image captioner in order to additionally enforce maximal key-frame diversity from the perspective of textual semantic content. Thus, the selected key-frames are encouraged to differ not only with regard to what objects they depict, but also with regard to their textual descriptions, which may additionally capture activities, scene context, etc. Empirical evaluation indicates that the proposed regularizer leads to state-of-the-art performance.

***Index Terms*—** key-frame extraction, Deep Neural Network, Long Short-Term Memory, Generative Adversarial Network, image captioning

1. INTRODUCTION

Automated video summarization consists in deriving succinct *summaries* of original, full-length videos, which capture the most important segments of the full input and jointly convey its essence in a compact manner. In *static* and *dynamic summarization*, the output is a set of still key-frames [1, 2] and a short trailer/skim [3, 4], respectively. In both cases, the goal is to select an informative, representative and temporally ordered subset of the original/full video, so that the remaining content can be discarded.

Initial unsupervised approaches to key-frame extraction involved clustering [5] or dictionary learning-based methods [6, 7]. A good summary is characterized by two main properties; its *saliency* and its *representativeness*. The former property suggests that the selected key-frames should be visually and/or semantically distinct from their temporal neighbours (local saliency) and/or from the rest of the video (global saliency, or *diversity*). The latter property implies that the selected key-frame set is capable of visually reconstructing the original/full video content.

Modern Deep Neural Network (DNN)-based supervised video summarization methods [8] typically rely on pretrained Convolutional Neural Networks (CNNs) to extract from each raw RGB video frame semantic representations, that describe visible scene objects. Each such representation is then fed to the summarization DNN, which selects the key-frames. Supervised training comes with high labor costs associated with required label annotation, places emphasis on subjective "ground-truth" and may lead to low generalization ability.

Unsupervised DNN-based video summarization promises a better solution. E.g., the adversarial reconstruction framework [9] is composed of two Long Short-Term Memory (LSTM) subnetworks: the *Summarizer* and the *Discriminator*. The first one consists of a *Selector* and an *Autoencoder* LSTM subnetwork, with the Autoencoder comprising an *Encoder* and a *Decoder*), sequentially arranged. From a functional standpoint, the Selector estimates a scalar importance score that expresses the suitability of each video frame to be included in the summary/key-frame set. Thus, the Autoencoder attempts to recreate the full input video sequence, given the selected key-frames, while the adversarially trained Discriminator classification module [10] tries to discern between summary-based reconstructions and original videos. After the overall architecture has been trained in a unified manner, all modules can be discarded except the Selector LSTM which is required for inference.

This basic framework focuses on the reconstructive ability of the summary, but [9] also included a Determinantal Point Process (DPP) regularizer [11] in the training process, pushing towards increased key-frame set global saliency/diversity. This diversity pertains mainly to the semantic content visible in the selected key-frames, since the Summarizer acts on convolutional representations of the original video frames and not on their raw RGB form. Thus, the visual DPP loss pushes towards summaries composed of key-frames that depict different objects. In essence, it operates by quantifying the representational variance of the video frames and penalizing candidate key-frame sets that do not capture significant percentage

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911 (AI4Media).

of the original video variance.

The adversarial reconstruction framework was subsequently improved upon in [12], [13], etc. However, despite significant progress, all unsupervised DNN-based methods that emphasize summary diversity focus only on the visible semantic content of the selected RGB key-frames, i.e., the scene objects, through a regularizer acting on the per-frame CNN-derived representations. To the best of our knowledge, *no effort has been expended towards enforcing key-frame set diversity with respect to different modalities under the adversarial reconstruction paradigm*.

This paper presents and evaluates a novel form of the DPP regularizer, readily applicable to the adversarial reconstruction framework for unsupervised key-frame extraction, which acts on video frame representations derived from a DNN-based image captioner. Thus, *the computed video summary is forced to be diverse not only with respect to the objects visible in its key-frames, but also with respect to the textual descriptions of these selected key-frames*. This regularizer can simply be added to the pool of employed loss functions for training the summarization DNN, assuming a pretrained DNN-based image captioner is available. Quantitative evaluation according to common protocols on two public, typically used benchmark datasets (TVSum, SumMe) indicates favourable results and non-negligible gains compared to baseline.

2. AUGMENTING SUMMARY DIVERSITY

Image captioning consists in generating a textual, natural-language description for a given RGB image. The primary challenge lies in two aspects: extracting adequate information from the visual content and generating grammatically correct, human-readable sentences. Several supervised DNN-based image captioning approaches exist, mostly involving architectures relying on CNNs and LSTMs.

The proposed method, called *DPP-caption loss* ($\mathcal{L}_{dpp-c}$), is a novel reformulation of the original *visual DPP loss* term ($\mathcal{L}_{dpp-v}$), which has been successfully applied as a regularizer for enforcing summary diversity in [9]. It relies on a pretrained DNN-based image captioner. At each iteration of the summarization DNN training, $\mathcal{L}_{dpp-c}$ pushes towards selecting key-frames that differ in their textual description according to the respective captioner output. *This enforces additional diversity in the derived summary, based on a non-object-centric semantic modality. For instance, an image caption may focus on depicted activities or scene context, instead of the visible objects.*

Below, the baseline adversarial reconstruction framework for unsupervised key-frame extraction is first detailed, before expanding upon the proposed novel regularizer.

2.1. Baseline Adversarial Reconstruction Framework

The basic adversarial reconstruction framework [9, 13] is detailed below.

Let $\mathbf{X} \in \mathbb{R}^{M \times N}$ be a video data matrix, where column $\mathbf{x}_i \in \mathbb{R}^M$ is the feature vector describing the i-th video frame, $1 \leq i \leq N$. Such a feature vector is typically a latent convolutional representation derived from a CNN R that has been pretrained for whole-image classification. Then, the baseline summarization architecture includes:

- An LSTM-based *Frame Selector* S parameterized by weights $\mathbf{w}_s$.
- An LSTM-based *Encoder* E parameterized by weights $\mathbf{w}_e$.
- An LSTM-based *Decoder* D parameterized by weights $\mathbf{w}_d$.
- An LSTM-based *Classifier* C parameterized by weights $\mathbf{w}_c$.

S, E and D jointly constitute the so-called *Summarizer*, with E and D being the two consecutive parts of an LSTM autoencoder. C acts as the *Discriminator* under a GAN framework. The entire architecture is trained end-to-end in an unsupervised manner. E, D and C are discarded after training is complete and only the optimized Frame Selector S is retained for inference/testing on unknown videos.

The forward pass of S is unfolded across N time instances. At the i-th time instance, S is fed $\mathbf{x}_i$ as input and outputs a corresponding scalar importance factor $s_i \in [0, 1]$. All s_i can be grouped in $\mathbf{s} \in \mathbb{R}^N$. The product $s_i \mathbf{x}_i$ is fed to E and this is performed consecutively for all i, resulting in a final LSTM hidden state vector $\mathbf{e} \in \mathbb{R}^H$ encoding the entire summary. Subsequently, $\mathbf{e}$ is fed to D which attempts to reconstruct the original $\mathbf{X}$, by outputting a reconstructed $\hat{\mathbf{x}}_i \in \mathbb{R}^M$, $1 \leq i \leq N$. Finally, the video reconstruction $\hat{\mathbf{X}}$ is forwarded to the Discriminator C as a "fake" training example, while the original video $\mathbf{X}$ is used as a "real" training example.

The following loss functions are employed during training:

- Reconstruction loss $\mathcal{L}_{recon} = \|\phi(\mathbf{X}) - \phi(\hat{\mathbf{X}})\|_2^2$, where $\phi(\mathbf{X})$ is the last hidden LSTM state of C when it is fed $\mathbf{X}$ as input and $\phi(\hat{\mathbf{X}})$ the corresponding hidden LSTM state when C is fed $\hat{\mathbf{X}}$. $\mathcal{L}_{recon}$ is used to update $\mathbf{w}_s$, $\mathbf{w}_e$ and $\mathbf{w}_d$.
- Original video loss $\mathcal{L}_{orig} = \left(1 - C(\mathbf{X})\right)^2$, which is the MSE between the original video label (i.e., 1) and the computed probability when C is fed $\mathbf{X}$ as input. $\mathcal{L}_{orig}$ is used to update $\mathbf{w}_c$.
- Summary loss $\mathcal{L}_{sum} = \left(C(\hat{\mathbf{X}})\right)^2$, which is the MSE between the summary label (i.e., 0) and the computed probability when C is fed $\hat{\mathbf{X}}$ as input. $\mathcal{L}_{sum}$ is used to update $\mathbf{w}_c$.
- Generator loss $\mathcal{L}_{gen} = \left(1 - C(\hat{\mathbf{X}})\right)^2$, which is the MSE between the original video label (i.e., 1) and the computed probability when C is fed $\hat{\mathbf{X}}$ as input. $\mathcal{L}_{gen}$ is used to update $\mathbf{w}_d$.

- Sparsity Loss $\mathcal{L}_{sparsity} = ||\frac{1}{N}\sum_{t=1}^{N} s_t - \sigma||_2$, which pushes the Selector towards assigning high importance (i.e., key-frame status probability) to a specific percentage of the total number of original video frames, defined by a scalar hyperparameter $\sigma \in [0, 1]$. This penalty term updates $\mathbf{w}_s$.

- Determinantal Point Process (DPP) loss $\mathcal{L}_{dpp-v} = -log\left(\frac{det\mathbf{L}(\mathbf{s})}{det(\mathbf{L}+\mathbf{I})}\right)$, where $\mathbf{L} \in \mathbb{R}^{N\times N}$ is a similarity matrix between every two hidden states of E and $\mathbf{L}(\mathbf{s})$ is a smaller square matrix cut down from $\mathbf{L}$ given $\mathbf{s}$ (which directly selects the summary/key-frame set). $\mathcal{L}_{dpp-v}$ is a diversity-inducing regularizer used to update $\mathbf{w}_s$.

2.2. DPP-caption Loss

The proposed novel regularizer $\mathcal{L}_{dpp-c}$ requires an LSTM-based image captioner, pretrained on a generic mass-scale annotated dataset, which we denote by P.

While training an unsupervised summarization DNN falling under the adversarial reconstruction framework, each video frame is forwarded to P (in inference mode), in parallel to feeding it to the Encoder E. Thus, the final hidden state of P encodes features representing a semantic textual description of said image, including visible objects, activities and scene context.

Then, $\mathcal{L}_{dpp-c}$ can be computed as a loss term in the following manner:

$$\mathcal{L}_{dpp-c} = -log\left(\frac{det\mathbf{P}(\mathbf{s})}{det(\mathbf{P}+\mathbf{I})}\right), \quad (1)$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ is a similarity matrix between every two final hidden states of the LSTM in P and $\mathbf{P}(\mathbf{s})$ is a smaller square matrix cut down from $\mathbf{P}$ given $\mathbf{s}$. $\mathcal{L}_{dpp-c}$ is also used to update $\mathbf{w}_s$.

Evidently, $\mathcal{L}_{dpp-c}$ induces a different kind of semantically informed diversity into the computed summary, in comparison to the original $\mathcal{L}_{dpp-v}$. This is because P encodes textual features capturing semantic properties (e.g., visible activities or scene context) that are potentially complementary to the visual features computed by R, which only represent scene objects. From a practical standpoint, the proposed method simply consists in adding $\mathcal{L}_{dpp-c}$ to the pool of the employed loss terms while training the complete summarization DNN model. After training is finished, P may be completely removed from the architecture; thus there is zero runtime overhead in inference mode.

3. EVALUATION

In order to evaluate the proposed method, the implementation [13] of the adversarial reconstruction framework (SUM-GAN-AAE) was adopted as a baseline. The reason behind this choice was solely practical; *in principle, the proposed method can be used to augment any other variant of the general framework, as well*. However, [13] does not include $\mathcal{L}_{dpp-v}$, thus the visual DPP loss term was implemented from scratch.

Method	TVSum	SumMe
SUM-FCN$_{unsupervised}$[14]	52.7%	41.5%
DR-DSN [15]	57.6%	41.4%
EDSN [16]	57.6%	42.6%
Unpaired VSN [17]	55.6%	47.5%
PCDL [18]	58.4%	42.7%
SUM-GAN-sl [12]	58.4%	47.8%
Cycle-SUM [19]	57.6%	41.9%
ACGAN [20]	58.5%	46.0%
[13]	58.3%	48.9%
[13] + $\mathcal{L}_{dpp-v}$	61.0%	56.5%
Proposed-A ([13] + $\mathcal{L}_{dpp-c}$)	62.6%	56.9%
Proposed-B ([13] + $\mathcal{L}_{dpp-v}$ + $\mathcal{L}_{dpp-c}$)	**63.5%**	**58.8%**

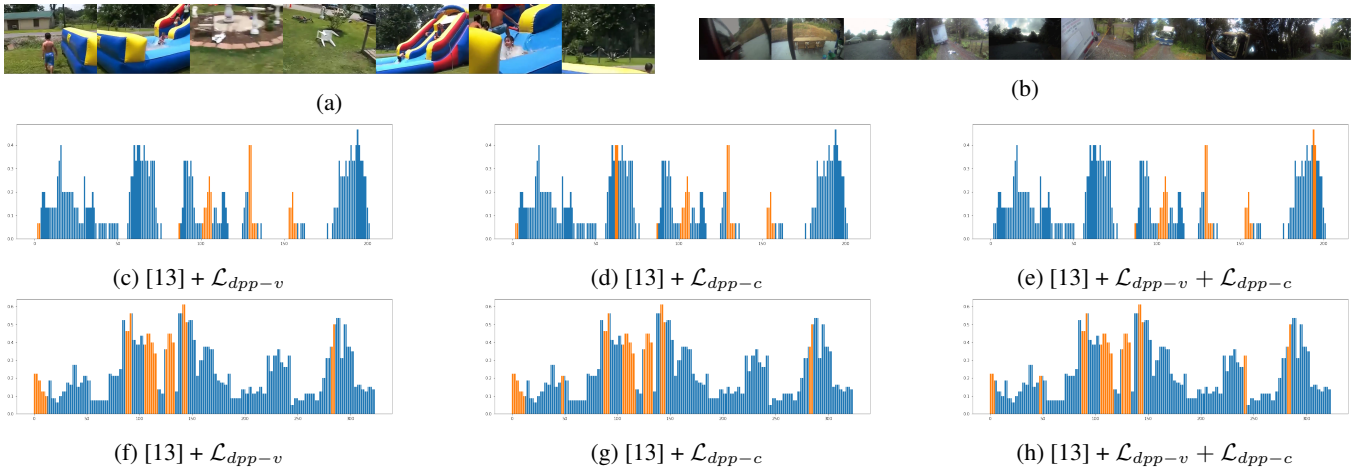
Table 1: Comparative evaluation of deep unsupervised key-frame extraction methods, using the F-score metric. Best results are in bold.

The employed image captioner P employed a typical Encoder-Decoder architecture. The Encoder was a ResNet-152 CNN, pretrained for whole-image classification on the generic ImageNet dataset. The CNN produces a 2048-dimensional vector representation capturing the semantic, object-centric content of the input image. Subsequently, this is fed to the LSTM Decoder, in order to predict a textual, natural-language caption for the given image. The LSTM has a 512-dimensional hidden state and is temporally unfolded for K time instances, where K is the maximum caption length (in words).

This image captioner was pretrained for 120 epochs in the generic COCO dataset, using the cross-entropy loss. Two Adam optimisers were employed and the respective learning rate values for the Encoder and the Decoder were 10^{-4} and $4\cdot10^{-4}$. During inference, the final hidden state of the LSTM Decoder encodes features of a textual image representation that comprises the corresponding caption. Thus it may incorporate complementary semantic information concerning not only the visible objects, but also the scene context, the visible activities, etc.

Typical quantitative evaluation protocols were followed [23]: sequences were subsampled at 2 FPS, video frame representations were extracted from the pool5 layer of a pretrained GoogLeNet CNN, serving as R, while all LSTM hidden state vectors were 500-dimensional. Training proceeded for 50 epochs, using Adam optimization and a learning rate of 10^{-4}. To infer the summary subset indices from the importance scores outputted by the Selector LSTM, a Knapsack algorithm was used [24] to temporally segment the video into subshots. Based on their importance, the subshots are sorted and, finally, the key-frame indices are selected.

Evaluation was conducted on two common public datasets:



(a) (b)

(c) [13] + $\mathcal{L}_{dpp-v}$ (d) [13] + $\mathcal{L}_{dpp-c}$ (e) [13] + $\mathcal{L}_{dpp-v} + \mathcal{L}_{dpp-c}$

(f) [13] + $\mathcal{L}_{dpp-v}$ (g) [13] + $\mathcal{L}_{dpp-c}$ (h) [13] + $\mathcal{L}_{dpp-v} + \mathcal{L}_{dpp-c}$

Fig. 1: 1a, 1b: Sampled video frames from test sequences 16 and 14 of SumMe [21] and TVSum [22], respectively. 1c, 1d, 1e: comparative evaluation of the summaries extracted from test video 16 (SumMe). 1f, 1g, 1h: comparative evaluation of the summaries extracted from test video 14 (TVSum). The height of each bar represents the video frame's respective ground-truth score, while the horizontal axis is the timeline. The orange bars represent the selected key-frames.

TVSum [22] and SumMe [21]. Each one was partitioned into 5 random splits, using a 80%-to-20% ratio for training and testing, respectively. The typically used F-Score metric was employed for evaluation. Table 1 depicts F-Score results for several recent DNN-based unsupervised key-frame extraction methods, given the common sparsity percentage of $\sigma = 15\%$. In all cases, the reported final figure is the mean F-Score performance across the 5 validation set splits. Since our implementation baseline [13] did not originally include the visual DPP loss term $\mathcal{L}_{dpp-v}$, the proposed method was evaluated both with ("Proposed-B") and without ("Proposed-A") $\mathcal{L}_{dpp-v}$ during training.

Evidently, augmenting the baseline codebase of [13] with the proposed $\mathcal{L}_{dpp-c}$ during training gives rise to non-negligible F-score gains and state-of-the-art performance at the inference stage. $\mathcal{L}_{dpp-c}$ alone (Proposed-A) leads to slightly better results than $\mathcal{L}_{dpp-v}$ alone, while using both regularizers (Proposed-B) yields the overall best performance compared to the best baseline [13] + $\mathcal{L}_{dpp-v}$: F-score gains of $+2.5\%/+2.3\%$ in TVSum/SumMe, respectively. Moreover, Figure 1 shows indicative qualitative results, comparing the behaviour of the best baseline ([13] + $\mathcal{L}_{dpp-v}$), Proposed-A ([13] + $\mathcal{L}_{dpp-c}$) and Proposed-B ([13] + $\mathcal{L}_{dpp-v}$ + $\mathcal{L}_{dpp-c}$) in two test sequences. It depicts a plot of the ground-truth video frame importance scores over time, while the orange bars represent the selected key-frames. Evidently, the proposed method leads to key-frames that are more concentrated around high-importance video segments.

4. CONCLUSIONS

This paper presented a novel, image captioning-based reformulation of the DPP regularizer, for unsupervised video summarization/key-frame extraction relying on the state-of-the-art adversarial reconstruction framework. Training with this regularizer augments summary diversity by pushing towards selecting key-frames that differ not only with regard to what objects they depict, but also with regard to their textual descriptions. This is because an image caption may also focus on depicted activities or scene context, along with the visible objects. The proposed $\mathcal{L}_{dpp-c}$ regularizer may be added to the pool of loss terms of any variant of the adversarial reconstruction framework, while only requiring a pretrained LSTM-based image captioner as a prerequisite. This can be discarded during inference, therefore the method induces zero runtime overhead after training has finished. Quantitative evaluation according to common protocols on two public, typically used benchmark datasets (TVSum, SumMe) showed favourable results and significant gains compared to baseline.

5. REFERENCES

[1] I. Mademlis, A. Tefas, and I. Pitas, "Regularized SVD-based video frame saliency for unsupervised activity video summarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[2] I. Mademlis, A. Tefas, and I. Pitas, "Summarization of human activity videos using a salient dictionary," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017.

[3] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Movie shot selection preserving narrative properties," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2016.

[4] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multimodal stereoscopic movie summarization conforming to narrative characteristics," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5828–5840, 2016.

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

[6] I. Mademlis, A. Tefas, and I. Pitas, "A salient dictionary learning framework for activity video summarization via key-frame extraction," *Information Sciences*, vol. 432, pp. 319–331, 2018.

[7] I. Mademlis, A. Tefas, and I. Pitas, "Greedy salient dictionary learning with optimal point reconstruction for activity video summarization," in *Proceedings of the International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018.

[8] B. Zhao, X. Li, and X. Lu, "TTH-RNN: Tensor-train hierarchical recurrent neural network for video summarization," *IEEE Transactions on Industrial Electronics*, 2020.

[9] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y.A. Bengio, "Generative adversarial nets," *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014.

[11] A. Kulesza and B. Taskar, "Determinantal Point Processes for machine learning," *arXiv preprint arXiv:1207.6083*, 2012.

[12] 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 *Proceedings of the International Workshop on AI for Smart TV Content Production, Access and Delivery*, 2019.

[13] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Unsupervised video summarization via attention-driven adversarial learning," in *Proceedings of the International Conference on Multimedia Modeling (MMM)*. Springer, 2020.

[14] M. Rochan, L. Ye, and Y. Wang, "Video summarization using Fully Convolutional sequence Networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[15] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[16] N. Gonuguntla, B. Mandal, and NB Puhan, "Enhanced deep video summarization network," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[17] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] B. Zhao, X. Li, and X. Lu, "Property-constrained dual learning for video summarization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 3989–4000, 2019.

[19] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

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

[21] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014.

[22] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[23] M. Kaseris, I. Mademlis, and I. Pitas, "Adversarial unsupervised video summarization augmented with dictionary loss," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2021.

[24] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014.