IEEE ICIP 2023. Author's accepted version. The final publication is available at https://ieeexplore.ieee.org/

SELECTING A DIVERSE SET OF AESTHETICALLY-PLEASING AND REPRESENTATIVE VIDEO THUMBNAILS USING REINFORCEMENT LEARNING

Evlampios Apostolidis[1,2] *Georgios Balaouras*[1] *Vasileios Mezaris*[1] *Ioannis Patras*[2]

[1] Information Technologies Institute (ITI), CERTH, Thermi 57001, Greece
[2] Queen Mary University of London, Mile End Campus, London E14NS, UK
{apostolid, mpalaourg, bmezaris}@iti.gr, i.patras@qmul.ac.uk

ABSTRACT

This paper presents a new reinforcement-based method for video thumbnail selection (called RL-DiVTS), that relies on estimates of the aesthetic quality, representativeness and visual diversity of a small set of selected frames, made with the help of tailored reward functions. The proposed method integrates a novel diversity-aware Frame Picking mechanism that performs a sequential frame selection and applies a reweighting process to demote frames that are visually-similar to the already selected ones. Experiments on two benchmark datasets (OVP and YouTube), using the top-3 matching evaluation protocol, show the competitiveness of RL-DiVTS against other SoA video thumbnail selection and summarization approaches from the literature.

Index Terms— Video thumbnail selection, reinforcement learning, aesthetic quality, representativeness, diversity

1. INTRODUCTION

Over the last years there is a tremendous growth of videos over the Web. To facilitate users' navigation in data collections, most video sharing platforms and social networks represent each video, in their data browsing interfaces, using one or a few thumbnails. However, manually selecting good thumbnails is a tedious and time-consuming process, as it requires a careful inspection of the entire content by a human editor. To accelerate this process, several methods have been proposed over the last years. Early approaches were based on rules about the optimal video thumbnail and extracted low-level (e.g., luminance) and mid-level features (e.g., appearance of faces) to assess frames' alignment with these rules [1, 2, 3]. More recent methods focused on specific characteristics of the video frames, such as their representativeness and aesthetic quality, and were based either on traditional feature extraction and clustering algorithms [4, 5, 6], or on the use of deep network architectures [7, 8, 9]. Finally, a few multimodal approaches take into account the users' intentions, expressed as textual queries [10, 11, 12].

Contrary to existing approaches that use similar thumbnail selection criteria [7, 9, 4], we propose a new method that considers also the frames' diversity during the selection and evaluation of video thumbnails. Moreover, instead of assessing frames' representativeness using Autoencoders [7], Generative Adversarial Networks (GANs) [9], or data clustering algorithms [4], our method uses a tailored reward function. Finally, the proposed method is the first to learn the video thumbnail selection task based on reinforcement learning and a set of reward functions. Our contributions are as follows:

- We introduce the use of reinforcement learning to learn video thumbnail selection based on rewards about the frames' aesthetics, representativeness and diversity.
- We design a new Frame Picking mechanism that takes into account the frames' diversity and discourages the selection of visually-similar frames.
- We integrate the designed Frame Picking mechanism in a novel network architecture, that learns how to select a diverse set of aesthetically-pleasing and representative video thumbnails, based on reinforcement learning.

2. RELATED WORK

In this section we focus on visual-based approaches as these are more closely related to the proposed method. Early works relied on hand-crafted rules on what indicates a good video thumbnail, and tailored features to assess frames based on these rules. Lian et al. [1] considered the appearance of faces, the variance of luminance and color diversity. Zhang et al. [2] took into account the frames' blurriness, visual saliency and pair-wise similarities. Choi et al. [3] defined cost functions to penalize frames with restricted appearance of faces/objects and blurred/shaky content. Song et al. [4] used low-level features (e.g. luminance, sharpness) to filter-out low-quality frames, and assessed the representativeness and aesthetics of the remaining ones based on data clustering and a stillness value, respectively. Tsao et al. [5] estimated the frames' attractiveness based on low- (e.g. sharpness, saturation) and high-level factors (presence of subtitles/persons). However, defining a complete set of rules for selecting good

This work was supported by the EU Horizon 2020 programme under grant agreement H2020-951911 AI4Media.

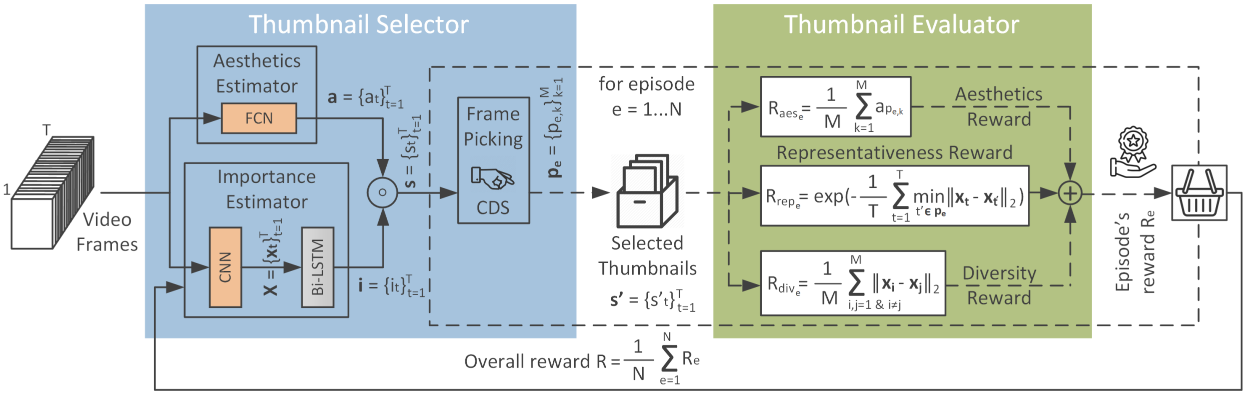
Thumbnail Selector
Aesthetics Estimator
FCN
Importance Estimator
CNN
Bi-LSTM
$\mathbf{X} = \{\mathbf{x_t}\}_{t=1}^{T}$
$\mathbf{a} = \{a_t\}_{t=1}^{T}$
$\mathbf{i} = \{i_t\}_{t=1}^{T}$
$\mathbf{s} = \{s_t\}_{t=1}^{T}$
$\circ$
T
1
Video Frames
Frame Picking
CDS
$\mathbf{p_e} = \{p_{e,k}\}_{k=1}^{M}$
for episode e = 1...N
Selected Thumbnails
$\mathbf{s'} = \{s'_t\}_{t=1}^{T}$
Thumbnail Evaluator
$R_{aes_e} = \frac{1}{M}\sum_{k=1}^{M} a_{p_{e,k}}$
Aesthetics Reward
Representativeness Reward
$R_{rep_e} = \exp(-\frac{1}{T}\sum_{t=1}^{T} \min_{t' \in \mathbf{p_e}} \|\mathbf{x_t} - \mathbf{x_{t'}}\|_2)$
$R_{div_e} = \frac{1}{M}\sum_{i,j=1\ \&\ i \neq j}^{M} \|\mathbf{x_i} - \mathbf{x_j}\|_2$
Diversity Reward
$\oplus$
Episode's reward R_e
Overall reward $R = \frac{1}{N}\sum_{e=1}^{N} R_e$

Fig. 1. The RL-DiVTS network architecture. Orange boxes indicate pretrained components and gray boxes denote trainable components and white boxes correspond to reward functions. Dashed lines represent iterative processes during a training epoch.

video thumbnails and extracting features for evaluating the frames against these rules, is a complex task.

To overcome the above shortcoming, recent works focused on a few commonly-desired features of video thumbnails, and examined the learning efficiency of deep networks. Gu et al. [7] assessed the frames' aesthetic quality using a CNN estimator pretrained on the AVA dataset [13], and evaluated their representativeness using a trainable Autoencoder and a thumbnail-to-video reconstruction process. Arthurs et al. [14] trained a variation of AlexNet [15] for classifying frames into good and bad thumbnails, and showed that adaptations of modern CNN classifiers can exhibit human-level performance on the aforementioned classification task. Pretorious et al. [16] performed a more extensive comparison of various CNNs for video thumbnail selection, using thumbnails of movies and TV series. Ren et al. [17] trained a Siamese CNN using annotations about the frames' ranking based on their representativeness and considering facial-related features. Finally, in [9] we utilized an LSTM-based adversarially-trained discriminator to measure the representativeness of the selected thumbnails, and combined its feedback with estimates about the thumbnails' aesthetics. However, these works rely on costly ground-truth data [14, 16, 17] or computationally-demanding network architectures [7, 9].

3. PROPOSED APPROACH

Network architecture: An overview of the RL-DiVTS network architecture is shown in Fig. 1. Given a video of T frames, at training time the Thumbnail Selector assesses the aesthetic quality and importance of each frame with the help of two estimators. The Aesthetic Estimator is a Fully Convolutional Network (FCN) proposed in [18], trained on the AVA dataset [13]. The assessment is done on a per frame basis and results in a sequence of scores that quantify the aesthetic quality of each video frame ($\boldsymbol{a} = \{a_t\}_{t=1}^{T}$ with $a_t \in [0, 1]$). The evaluation of the frames' importance is performed by modeling their temporal dependence. The Importance Estimator extracts one feature vector per frame using the pool5 layer of a model of GoogleNet [19] trained on ImageNet [15], and passes the extracted feature vectors ($\boldsymbol{X} = \{\boldsymbol{x_t}\}_{t=1}^{T}$) to a bi-directional LSTM that models the frames' temporal dependence and assigns a score to each frame that represents its importance ($\boldsymbol{i} = \{i_t\}_{t=1}^{T}$ with $i_t \in [0, 1]$). The computed scores about the frames' aesthetic quality and importance are then fused via their Hadamard product (denoted as $\circ$ in Fig. 1), resulting to a new sequence of scores ($\boldsymbol{s} = \{s_t\}_{t=1}^{T}$) that is used by the Frame Picking mechanism.

To promote the selection of diverse frames, we introduce a Categorical Distribution Sampler (CDS) that selects frames sequentially by sampling from an appropriate distribution. At the first step, this distribution is based on $\boldsymbol{f_1} = \{f_t\}_{t=1}^{T}$ (computed as $\boldsymbol{f_1} = N(\boldsymbol{s})$, where $N()$ denotes min-max normalization) and the sampling process results in the first picked frame (p_1) and a log probability of picking this sample from the distribution (lp_1). At each subsequent step m (with $m \in [2, M]$), this distribution is based on $\boldsymbol{f_m} = N(\boldsymbol{f_{m-1}} \circ (1 - \boldsymbol{u_{p_{m-1}}}))$, where $\boldsymbol{u_{p_{m-1}}}$ denotes the row of the frames' (cosine) similarity matrix that corresponds to the picked frame at step $m - 1$ (see Fig. 2) and the Hadamard product within $N()$ effects a re-weighting, i.e., demotes the selection of frames that are visually-similar to the already picked ones. After the end of the M steps the Frame Picking mechanism defines a set of picked frames $[p_1, ..., p_M]$ and a set of log probabilities $[lp_1, ..., lp_M]$; the latter are used to compute the expected reward in the context of episodic reinforcement learning.

The output of the frame selection process for the e^{th} episode (see $\boldsymbol{p_e} = \{p_{e,k}\}_{k=1}^{M}$ in Fig. 1) is assessed by the Thumbnail Evaluator, in terms of aesthetic quality, representativeness and diversity, using the reward functions in Eq. 1, 2 (proposed in [20]) and 3, respectively. The overall reward for the current episode is then formed by the weighted sum in Eq. 4 (denoted as $\oplus$ in Fig. 1), where D projects R_{rep_e} in the same scale with the other rewards. Finally, the average reward across all the N episodes is the feedback of the Thumbnail Evaluator for the current training sample.

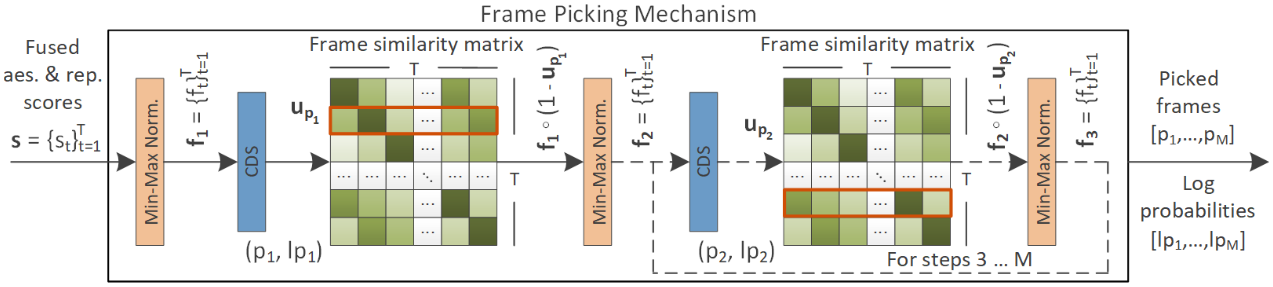

Fig. 2. Processing steps of the proposed Frame Picking mechanism. Dashed lines indicate iterative processes during an episode.

$$R_{aes_e} = \frac{1}{M}\sum_{k=1}^{M} a_{p_{e,k}} \tag{1}$$

$$R_{rep_e} = exp(-\frac{1}{T}\sum_{t=1}^{T} \min_{t' \in \boldsymbol{p_e}} \|\boldsymbol{x_t} - \boldsymbol{x_{t'}}\|_2) \tag{2}$$

$$R_{div_e} = \frac{1}{M}\sum_{i,j=1}^{M} \|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2 \tag{3}$$

$$R_e = \alpha \cdot R_{aes_e} + \beta \cdot D \cdot R_{rep_e} + \gamma \cdot R_{div_e} \tag{4}$$

At inference time, only the Thumbnail Selector is used. Given a video of T frames, it estimates the frames' aesthetic quality and visual importance, and passes the sequence of fused scores $\boldsymbol{s}$ to the proposed Frame Picking mechanism. The latter makes M picks through the sequential process described above, leading to a sequence of frame-level scores ($\boldsymbol{s'} = \{s'_t\}_{t=1}^{T}$), as depicted in Fig. 1. In this sequence, the $T - M$ zero values indicate non-selected frames of the video, and the M non-zero values signify the suitability of each of the M selected frames to be a video thumbnail, based on the estimated aesthetic quality and importance of its visual content. Following, the k top-scoring frames (where k equals to 3 in our experiments) are selected as the video thumbnails.

Training strategy: For training RL-DiVTS we utilize the episodic REINFORCE algorithm [21]. In each episode, the Thumbnail Selector picks M frames following the processing steps depicted in Fig. 2. Then, the Thumbnail Evaluator computes the overall reward for the current episode, based on Eq. 4. To compute the expected reward, we normalize the sum of the log probabilites of the M sequential actions made by the Frame Picking mechanism based on the number of picks, and multiply the computed value with the overall reward after subtracting from the reward a constant baseline b which is computed as the moving average of the received rewards in the previous episodes (to avoid high variance in the computed gradients [22]). The training loss is formed so as to minimize the negative expected reward, and after the end of all training episodes, the gradients are computed based on the accumulated loss value. Based on this strategy the Thumbnail Selector learns a policy for scoring the video frames, by maximizing the expected rewards from the Thumbnail Evaluator.

4. EXPERIMENTS

Datasets and evaluation protocol: We assessed the performance of RL-DiVTS using the publicly-available datasets and evaluation protocol of [7]. The OVP dataset is composed of 50 videos (up to 3.5 min. long) with diverse content (e.g., documentaries, lecture videos). The YouTube dataset contains 50 videos (up to 9.5 min. long) of different types (e.g., news, TV-shows). Each video has been annotated by 5 users in the form of key-frames. As in [7], for each video we considered the 3 most selected key-frames among all annotators as its ground-truth thumbnails, and we estimated their similarity with the automatically-selected ones using the Structural Similarity Index (SSIM); we called it a match if SSIM score > 0.7. For evaluation, we applied the "top-3 matching" approach of [7], that measures the overlap between the top-3 machine- and human-selected thumbnails per video. We expressed this overlap as a scalar ranging in $[0, 1]$ and computed the average score over all videos of the test set. As a note, the adopted evaluation protocol is different and much more challenging from the one in [9], which declares a hit (i.e., takes value "1") if at least one of the top-3 machine-selected thumbnails matches one (or more) of the top-3 ground-truth ones, according to the same SSIM-related threshold.

Implementation details: All videos were downsampled to 2 fps. The Importance Estimator contains a 2-layer bi-directional LSTM with 512 hidden units, that is trained in a full-batch mode using the Adam optimizer. The number of selected frames $M = 6$ and the number of episodes $N = 10$. Factor D is set equal to $5 \cdot 10^3$ and $2.5 \cdot 10^3$ for OVP and YouTube respectively. In Eq. 4, $\alpha = 0.35$, $\beta = 0.35$ and $\gamma = 0.3$. Training runs for 150 epochs and we select the model that maximizes the overall reward on the training set. Following the paradigm of most SoA video summarization works [23], we split the used dataset into a training (containing 80% of data) and a testing (remaining 20% of data) set. From YouTube, we exclude 10 cartoon videos, as the used feature extraction (GoogleNet) and aesthetics estimation (FCN) components cannot provide meaningful representations and measurements for cartoon videos. To reduce the impact of the used data split and the network's initialization, we run our experiments using 5 different randomly-created

	OVP	YouTube
Baseline (Random)	8.63 ± 2.50	4.41 ± 1.77
AC-SUM-GAN [24]	7.87 ± 3.41	7.33 ± 0.70
CA-SUM [25]	7.60 ± 2.85	8.00 ± 3.56
Hecate-VTS [4]	11.72	16.47
ReconstSum [7]	12.18	**18.25**
ARL-VTS [9]	12.50 ± 3.37	7.83 ± 1.49
RL-DiVTS (proposed)	**25.33** ± 3.97	17.50 ± 2.57

Table 1. Performance comparison of RL-DiVTS with a baseline (random-picking) approach, and a set of SoA video thumbnail selection and summarization methods.

	Training time (sec/epoch)		# Param.
	OVP	YouTube	(in Millions)
ARL-VTS [9]	38.41	62.43	28.36
RL-DiVTS	2.33	2.70	12.60

Table 2. Comparison of RL-DiVTS and ARL-VTS, in terms of training time and amount of learnable parameters.

splits and 5 different random seeds, and we report the average performance and the standard deviation over these 25 runs. The experiments were carried out using an NVIDIA RTX 2080 Ti. The PyTorch implementation of RL-DiVTS is available at: https://github.com/e-apostolidis/RL-DiVTS.

Performance comparisons: We compared RL-DiVTS against a baseline that selects video thumbnails randomly, and a set of SoA video thumbnail selection and summarization methods from the literature. Following [7], we considered two video summarization methods ([24] and [25]) with public implementations. The results of this comparison are shown in Tab. 1. The scores for Hecate-VTS [4] and ReconstSum [7] are the ones reported in [7], as their reproduction was not feasible due to limited implementation details in [7]; e.g., there are no details about the used training/testing samples. These results show that RL-DiVTS performs consistently well on both datasets, being by far the top-performing one on OVP and the second best-performing one (slightly bellow the best one) on YouTube. Moreover, it is more suitable for thumbnail selection, compared to the examined summarization methods. Finally, compared to our previous ARL-VTS method [9], RL-DiVTS brings a noticeable performance improvement on both datasets. Moreover, it exhibits significant gains w.r.t. training time and memory footprint. The results in Tab. 2 demonstrate that replacing the GAN-based Representativeness Evaluator of ARL-VTS by a reward function, reduced the needed training time by more than 16 and 23 times for the OVP and YouTube videos, respectively. Moreover, this replacement removed the most computationally-demanding module of ARL-VTS, as indicated by the significantly reduced number of learnable parameters of RL-DiVTS.

Ablation study: To assess the contribution of each of the adopted video thumbnail evaluation criteria, we conducted an ablation study including three variants of RL-DiVTS that do not take into account one of these criteria. In each case, the overall reward is formed by averaging the two remaining rewards. To evaluate the impact of the proposed Frame Picking mechanism, we examined another variant of RL-DiVTS ("RL-DiVTS w/o CDS") that selects all M frames at once. The results in Tab. 3 show that considering the aesthetic quality is of major importance when selecting video thumbnails, as ignoring aesthetics leads to a big performance drop on both datasets. The frames' representativeness is also important, as excluding this aspect leads to consistently lower performance. In addition, taking into account the diversity of the selected frames during training is beneficial for the method's performance on YouTube, while it leads to similar levels of performance on OVP. Finally, incorporating knowledge about the frames' diversity during the frame-picking process positively affects the method's performance (especially on YouTube). Overall, the results in Tab. 3 indicate that the removal of either of the utilized criteria and the integrated Frame Picking mechanism results in a noticeable performance degradation in, at least, one of the used datasets. Finally, from Tab. 4 we see that picking fewer or more than 6 frames during training leads to reduced performance in both datasets.

	OVP	YouTube
RL-DiVTS w/o AES	14.13 ± 2.96	10.33 ± 1.73
RL-DiVTS w/o REP	20.53 ± 1.91	13.17 ± 1.09
RL-DiVTS w/o DIV	**26.40 ± 1.30**	14.33 ± 1.49
RL-DiVTS w/o CDS	24.67 ± 3.16	15.00 ± 1.44
RL-DiVTS (proposed)	25.33 ± 3.97	**17.50 ± 2.57**

Table 3. Ablation study w.r.t. the utilized rewards and the diversity-aware CDS-based Frame Picking mechanism.

Frames	OVP	YouTube
3	20.80 ± 1.66	13.67 ± 1.73
6	**25.33 ± 3.97**	**17.50 ± 2.57**
9	19.33 ± 1.94	11.67 ± 3.12

Table 4. Performance of RL-DiVTS when varying the amount of picked frames during training.

5. CONCLUSIONS

In this work we proposed the RL-DiVTS method for video thumbnail selection. A Thumbnail Selector estimates the frames' aesthetics and importance and integrates a diversity-aware Frame Picking mechanism. Then, a Thumbnail Evaluator assesses the aesthetic quality, representativeness and diversity of the selected frames using tailored reward functions. The overall reward is used to learn how to select a diverse set of aesthetically-pleasing and representative thumbnails based on reinforcement learning. Comparisons with SoA video thumbnail selection and summarization approaches showed the competitive performance of RL-DiVTS on two datasets.

6. REFERENCES

[1] H. Lian et al., "Automatic video thumbnail selection," in *2011 Int. Conf. on Multimedia Technology*, 2011, pp. 242–245.

[2] W. Zhang et al., "A novel framework for web video thumbnail generation," in *2012 Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, 2012, pp. 343–346.

[3] J. Choi et al., "A framework for automatic static and dynamic video thumbnail extraction," *Multimedia Tools and Applicat.*, vol. 75, no. 23, pp. 15975–15991, 2016.

[4] Y. Song et al., "To click or not to click: Automatic selection of beautiful thumbnails from videos," in *25th ACM Int. on Conf. on Information and Knowledge Management*. 2016, p. 659–668, ACM.

[5] C. Tsao et al., "Thumbnail image selection for VOD services," in *Proc. of the 2019 IEEE Conf. on Multimedia Information Processing and Retrieval*, 2019, pp. 54–59.

[6] Y. Chen et al., "Mobile media thumbnailing," in *Int. Conf. on Multimedia Retrieval (ICMR)*. 2015, p. 665–666, ACM.

[7] H. Gu et al., "From thumbnails to summaries - A single deep neural network to rule them all," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[8] B. Zhao et al., "Automatic generation of informative video thumbnail," in *8th Int. Conf. on Digital Home*, 2020, pp. 254–259.

[9] E. Apostolidis et al., "Combining adversarial and reinforcement learning for video thumbnail selection," in *Int. Conf. on Multimedia Retrieval (ICMR)*. 2021, p. 1–9, ACM.

[10] A. B. Vasudevan et al., "Query-adaptive video summarization via quality-aware relevance estimation," in *25th ACM Int. Conf. on Multimedia (ACM MM)*. 2017, p. 582–590, ACM.

[11] Y. Yuan et al., "Sentence specified dynamic video thumbnail generation," in *27th ACM Int. Conf. on Multimedia (ACM MM)*. 2019, p. 2332–2340, ACM.

[12] M. Rochan et al., "Sentence guided temporal modulation for dynamic video thumbnail generation," in *British Machine Vision Conf. (BMVC)*, 2020.

[13] N. Murray et al., "AVA: A large-scale database for aesthetic visual analysis," in *IEEE Conf. on Computer Vision and Patt. Recog. (CVPR)*, 2012, pp. 2408–2415.

[14] N. Arthurs et al., "Selecting YouTube video thumbnails via Convolutional Neural Networks," Tech. Rep., Stanford, 2017.

[15] A. Krizhevsky et al., "Imagenet classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[16] K. Pretorious et al., "A comparative study of classifiers for thumbnail selection," in *Int. Joint Conf. on Neural Networks*, 2020, pp. 1–7.

[17] J. Ren et al., "Best frame selection in a short video," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2020, pp. 3201–3210.

[18] K. Apostolidis et al., "Image aesthetics assessment using Fully Convolutional Neural Networks," in *25th Int. Conf. on MultiMedia Modeling (MMM)*. 2019, pp. 361–373, Springer International Publishing.

[19] C. Szegedy et al., "Going deeper with convolutions," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[20] K. Zhou et al., "Deep reinforcement learning for unsupervised video summarization with diversity - representativeness reward," in *AAAI Conf. on Artificial Intelligence (AAAI)*, 2018.

[21] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3–4, pp. 229–256, 1992.

[22] E. Greensmith et al., "Variance reduction techniques for gradient estimates in reinforcement learning," in *Advances in Neural Information Processing Systems*. 2001, vol. 14, MIT Press.

[23] E. Apostolidis et al., "Video summarization using deep neural networks: A survey," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1838–1863, 2021.

[24] E. Apostolidis et al., "AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for unsupervised video summarization," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3278–3292, 2021.

[25] E. Apostolidis et al., "Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames," in *Int. Conf. on Multimedia Retrieval (ICMR)*. 2022, p. 407–415, ACM.