Combining Adversarial and Reinforcement Learning for Video Thumbnail Selection

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

Eleni Adamantidou
CERTH-ITI
Thessaloniki, Greece, 57001
adamelen@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

This paper presents a new method for unsupervised video thumbnail selection. The developed network architecture selects video thumbnails based on two criteria: the representativeness and the aesthetic quality of their visual content. Training relies on a combination of adversarial and reinforcement learning. The former is used to train a discriminator, whose goal is to distinguish the original from a reconstructed version of the video based on a small set of candidate thumbnails. The discriminator's feedback is a measure of the representativeness of the selected thumbnails. This measure is combined with estimates about the aesthetic quality of the thumbnails (made using a SoA Fully Convolutional Network) to form a reward and train the thumbnail selector via reinforcement learning. Experiments on two datasets (OVP and Youtube) show the competitiveness of the proposed method against other SoA approaches. An ablation study with respect to the adopted thumbnail selection criteria documents the importance of considering the aesthetics, and the contribution of this information when used in combination with measures about the representativeness of the visual content.

CCS CONCEPTS

• **Computing methodologies** → **Video summarization**; **Machine learning algorithms**.

KEYWORDS

Video thumbnail selection, Deep neural networks, Generative adversarial networks, Reinforcement learning, Unsupervised learning

ACM Reference Format:
Evlampios Apostolidis, Eleni Adamantidou, Vasileios Mezaris, and Ioannis Patras. 2021. Combining Adversarial and Reinforcement Learning for Video Thumbnail Selection. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3460426.3463630

ICMR '21, August 21–24, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8463-6/21/08…$15.00
https://doi.org/10.1145/3460426.3463630

1 INTRODUCTION

In the last few years we are witnessing a constantly growing popularity of social networks and video sharing platforms, that was fueled - to a large extent - by a strong engagement of users with devices carrying video recording and online content sharing functionalities (e.g., smartphones, tablets and wearable cameras). This technological environment stimulated a tremendous growth of videos over the Internet. To facilitate users' navigation in endless collections of video content, most video sharing platforms and social networks represent each video, in their browsing interfaces or when displaying lists of search results, using a thumbnail. Given the plethora of online-available video content, the video thumbnail plays a key role in terms of content consumption as it significantly affects users when deciding whether to watch or skip a video.

Typically, a single key-frame is extracted from the video and used as thumbnail. To increase descriptiveness, some video sharing platforms (e.g., YouTube) provide a more vivid representation of the video content using animated GIFs (composed of a few key-frames) or short segments of the video. In any case, selecting a good thumbnail is a tedious and time-consuming process, as it requires a careful inspection of the entire content and a manual selection of one or more representative and aesthetically-pleasing key-frames. 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 [7, 13, 32]. More recent methods focused on a few characteristics that were identified as the most important ones for thumbnail selection, and relate to the representativeness and the aesthetic quality or attractiveness of the visual content. These methods are based either on traditional feature extraction and clustering algorithms [23, 25], or on the use of deep network architectures [9]. Finally, a few recent works associate video thumbnail selection with the users' intentions when searching for video content online, and propose multimodal approaches for dynamic video thumbnail selection, according to textual user queries [14, 17, 26, 30].

In this work we tackle video thumbnail selection as a video summarization task, where the goal is to select one or more key-frames that provide a representative and aesthetically-pleasing synopsis of the video content. In contrast to existing approaches that use the same thumbnail selection criteria [9, 23] (a more detailed comparison with these approaches, that highlights the novelty of our method, is given in the last paragraph of Section 2), we propose a new combination of adversarial and reinforcement learning to train a novel deep-learning architecture for video thumbnail selection. An adversarially-trained discriminator is used to make estimates about the representativeness of one or more selected key-frames of the video. These estimates are used in combination with measurements about the aesthetic quality of the visual content to form a reward signal. This reward signal is finally used to train the developed architecture in a fully-unsupervised manner, based on reinforcement learning. Our contributions are as follows:

- We introduce the use of reinforcement learning to learn the task of video thumbnail selection based on estimates about the representativeness and aesthetic quality of the visual content of the video frames.
- We propose a novel architecture and a training pipeline that combines the principles of adversarial and reinforcement learning. An adversarially-trained discriminator is used to measure the representativeness of the selected thumbnails, and its feedback is used in combination with estimates about the aesthetic quality to form a reward signal and train the video thumbnail selector via reinforcement learning.
- We conduct an ablation study that highlights the importance of considering the aesthetic quality of video frames when selecting a thumbnail, and shows the contribution of this type of information when used in combination with estimates about the thumbnail's representativeness, as proposed.

2 RELATED WORK

Several approaches were proposed over the last years to automate the video thumbnail selection process. In the sequel we focus on methods that rely solely on the visual content, as these are more closely related to the proposed approach. For the sake of completeness, though, we also briefly report on methods that exploit additional modalities of the video and/or auxiliary data.

Early visual-based approaches relied on hand-crafted rules about the optimal video thumbnail, and tailored features and mechanisms to assess video frames' alignment with these rules. In [13], video thumbnail selection is associated with the appearance of faces, the variance of luminance, and the color diversity. The extracted features are used by a fusion mechanism that computes a score indicating the appropriateness of a frame to be used as a thumbnail. In [32], video thumbnails are selected based on their visual quality, accessibility and thematic relevance. Visual quality is estimated by the degree of blurriness [6]. Accessibility is evaluated using a visual saliency model [10]. Thematic relevance is measured based on the pair-wise similarities of shot-level key-frames. A more recent approach [7] uses mid- and low-level features and a set of energy cost functions that penalize the selection of frames with: i) limited appearance or bad placement of faces/objects, ii) fast object movement, iii) blurred content, and iv) limited scene steadiness. The work of [23] is the first to correlate video thumbnail selection with the frames' visual quality. Initially, low-quality frames are discarded by examining luminance, sharpness and uniformity. Then, aesthetically-pleasing key-frames are extracted based on frame clustering and a stillness value. Finally, these key-frames are evaluated according to their relevance to the video content (quantified by the size of the cluster containing the key-frame), and their aesthetic quality (estimated using the stillness value or a trained random forest regression model [5]). The approach in [25] examines several low- and high-level factors (e.g., sharpness, saturation, brightness, and the presence of subtitles and faces) to filter-out non-attractive frames, and evaluates the remaining ones according to their representativeness using a clustering-based approach similar to [23]. The main shortcoming of the above discussed approaches resides in the fact that the definition of a complete set of commonly-accepted and content-independent rules about the optimal video thumbnail, as well as the engineering of the extracted features for evaluating the video frames against these rules, are both highly-complex tasks.

To overcome the aforementioned shortcoming, some recent works indicated a few commonly-desired characteristics for a video thumbnail, and tried to build thumbnail selection mechanisms by exploiting the learning efficiency of deep network architectures. The approach in [9] focuses on the representativeness and aesthetic quality of video frames. Building on the idea of [18], it assesses the representativeness of a sparse set of key-frames by using them to reconstruct the original video via an auto-encoding process. The aesthetic quality is evaluated using a CNN-based estimator pre-trained on the AVA dataset [19]. The sum of the computed scores about the aesthetics and importance of video frames is used to weight the video frames before the reconstruction process. Training is unsupervised and the goal is to learn how to select a set of representative and aesthetically-pleasing frames. The work of [4] examines the performance of two CNNs after being trained for classifying frames into good and bad thumbnails. Training is performed based on a set of Youtube videos and the assumption that the thumbnails of videos with more than 1 million views are good examples and thumbnails of videos with less than 100 views are bad examples. A more extensive comparison of various CNNs for video thumbnail selection is reported in [20]. Finally, [21] uses a set of CNNs to select the best frame from short frame sequences. In particular, a Siamese CNN is trained using pairs of images and a piece-wise ranking loss. Information about the appearance and quality of faces is incorporated using two additional CNNs. Their output is used during training the Siamese CNN, to learn a frame ranking policy that considers also the facial features of video frames.

A few other methods exploit information from additional modalities or auxiliary sources. An early approach [27] uses keywords from the textual video metadata to retrieve images from a database and select visually-similar frames of the video as thumbnails. [28] exploits textual metadata and audio to build a latent representation of the entire content and select a frame which is the nearest one to this representation in the learned latent space. The algorithm in [14] selects a frame that is representative of the video content and specific to the intent of the user's query using a dual cross-media relevance model [16]. The method in [17] utilizes a deep visual-semantic embedding model to form a latent space and estimate the relevance between the user's query and the video frames. Building on [17], the work in [26] presents a quality-aware relevance estimation model which can capture the query-independent frame-quality properties in the visual semantic embedding procedure. Finally, [30] describes a dynamic thumbnail selection process that relies on a temporal conditioned pointer network and a sentence-specified video graph convolutional network.

Proc. ACM ICMR 2021. Authors' accepted version. The final paper is avaiable at https://doi.org/10.1145/3460426.3463630

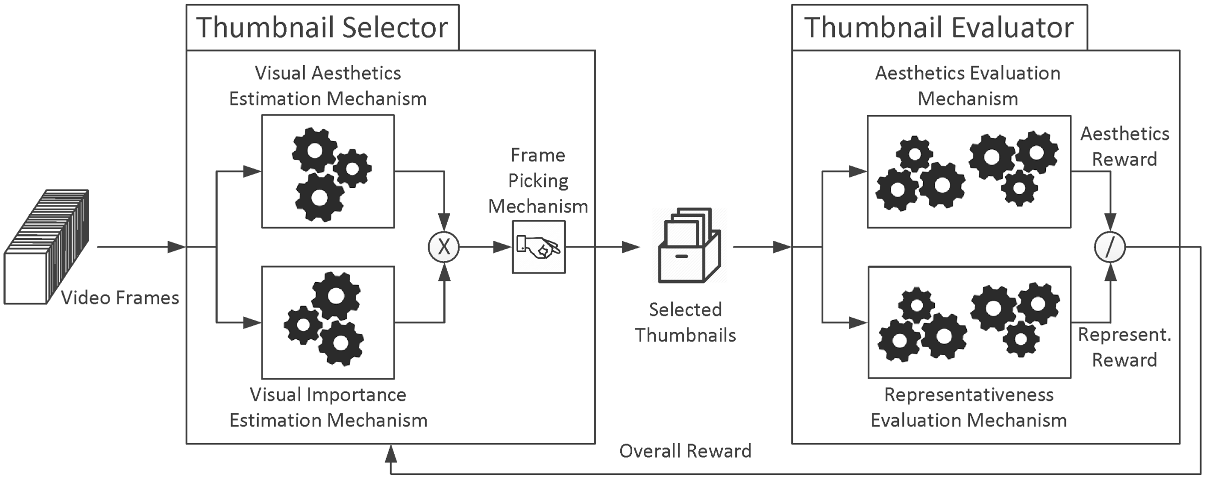
Thumbnail Selector
Visual Aesthetics Estimation Mechanism
Frame Picking Mechanism
X
Video Frames
Visual Importance Estimation Mechanism
Selected Thumbnails
Thumbnail Evaluator
Aesthetics Evaluation Mechanism
Aesthetics Reward
/
Represent. Reward
Representativeness Evaluation Mechanism
Overall Reward

Figure 1: Overview of the proposed approach for unsupervised learning of video thumbnail selection.

Based on the aforementioned literature review, in terms of the used modalities and criteria for video thumbnail selection, the proposed approach is most closely related to [23] and [9] that choose thumbnails based on measures about the aesthetic quality and representativeness of their visual content. However, contrary to [23], i) we assess the aesthetics using a SoA deep-learning-based approach (instead of using low-level features such as luminance and sharpness), and ii) we estimate representativeness using a trainable discriminator (instead of using a clustering algorithm). Moreover, in contrast to [9], i) we maximize the similarity between the original and the reconstructed version of the video based on the selected set of key-frames, using an adversarially-trained discriminator (instead of directly comparing them and trying to reduce a relevant loss), and ii) we utilize the computed estimates about the aesthetics also as part of a reward signal that is used to train our model via reinforcement learning (instead of using such estimates only as part of the frames' weighting before the video reconstruction process).

3 PROPOSED APPROACH

This section presents the proposed approach. It starts by discussing the main concept behind the design of the analysis pipeline (Sec. 3.1). Then, it describes the building blocks of the developed network architecture and the data flow at the training and inference time (Sec. 3.2). Finally, it explains the adopted strategy for training of the network (Sec. 3.3). With respect to the used notation: capital bold letters denote matrices, small bold letters denote vectors and non-bold letters (either capital or small) denote scalar values.

3.1 Overview

The overview of the proposed approach is depicted in Fig. 1. The sequence of video frames is given as input to the Thumbnail Selector. Each individual frame is evaluated by two internal mechanisms that make estimates about the aesthetics and importance of the visual content. The assessment of the aesthetic quality is performed on a per frame basis; i.e., to assess the quality of a given frame, the relevant mechanism examines the visual content of this particular frame only. The evaluation of the visual importance is performed by modeling the temporal dependencies of the entire frame sequence. The fused output of the aforementioned mechanisms - formed through an element-wise multiplication process that is represented by the $\otimes$ symbol in Fig. 1 - is then utilized by a frame picking mechanism, which selects a set of key-frames. The latter is based on a set of discrete sampled actions over a multinomial distribution that follows the distribution of the fused scores.

The set of selected key-frames is then forwarded to the Thumbnail Evaluator, and evaluated according to its aesthetic quality and representativeness. The former is measured as the mean of the computed aesthetic values for the selected key-frames. The latter is estimated by quantifying the similarity between the original video and a reconstructed version of it based on the set of selected key-frames. The general approach of using Generative Adversarial Networks to estimate this similarity was first proposed in [18] and further extended by several other SoA video summarization algorithms (e.g., [11, 12, 29]) as a means to assess the representativeness of a set of key-frames that will be eventually used to generate a static (a.k.a. video storyboard) or dynamic video summary (a.k.a. video skim). The fused output of the aforementioned assessments - computed by an average operator that is represented by the $\oslash$ symbol in Fig. 1 - forms the feedback of the Evaluator with respect to the representativeness and aesthetic quality of the set of candidate thumbnails. This feedback is utilized as a reward signal for training the Thumbnail Selector based on reinforcement learning.

Building on the above, we developed a network architecture (presented in Sec. 3.2) and a pipeline for its unsupervised training (described in Sec. 3.3). By combining adversarial and reinforcement learning, the Thumbnail Selector utilizes the received feedback from the Evaluator and, as the training proceeds, it progressively learns how to select a small set of video key-frames that provide a representative and aesthetically-pleasing synopsis of the video. The frame or frames with the highest values are selected as thumbnails.

3.2 Network architecture

The developed network architecture is shown in Fig. 2. To present each different component, in the sequel we describe the processing

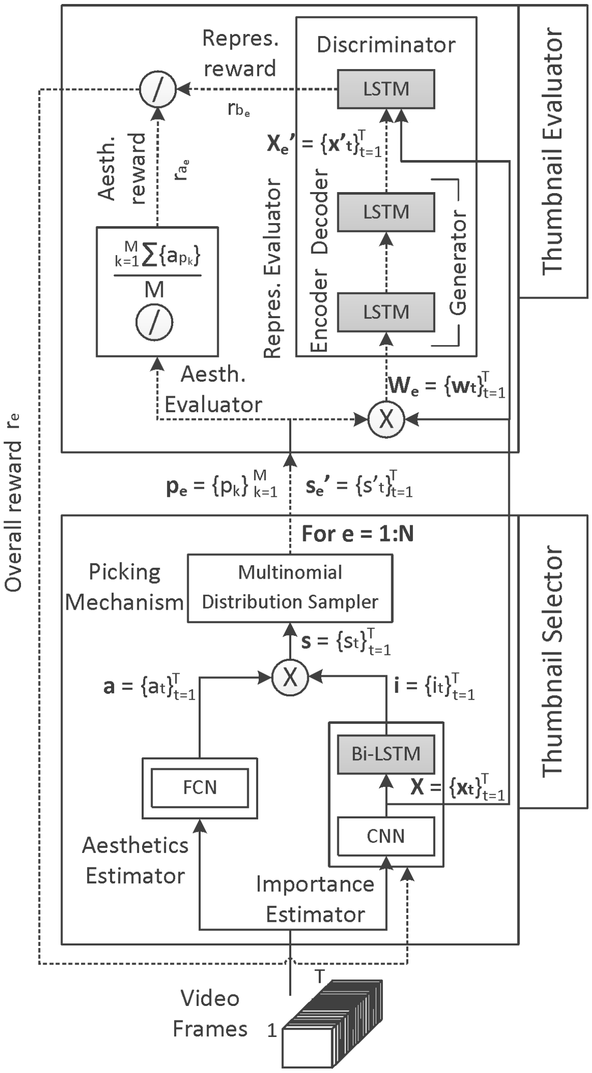

Figure 2: The proposed network architecture. Shaded boxes indicate trainable parts. Dashed lines represent iterative processes when training the Importance Estimator.

steps during training time where the entire network is utilized. At inference time, only the Thumbnail Selector is used; further details about this are given in the last paragraph of this section.

Given an input video of T frames, the Thumbnail Selector initially assesses the visual aesthetic quality and importance of each frame with the help of two estimators. The Aesthetic Estimator is a model of the Fully Convolutional Network (FCN) from [3], pretrained on the AVA dataset [19]. This dataset was built to assist the development and evaluation of image aesthetic quality assessment methods. However, we exploit the knowledge of a network trained on this dataset to assist the video thumbnail selection task, given the lack of a more relevant dataset, and similarly to [9]. The network from [3] is an extension of the VGG16 architecture [22], that utilizes skip connections and a setup for minimizing the sizing distortions of the input image. It exhibits SoA performance on the task of image aesthetics assessment[1]. This estimator gets as input the raw video and produces a sequence of scores that quantify the aesthetic quality of each video frame (frame-level scores $\boldsymbol{a} = \{a_t\}_{t=1}^T$ with $a_t \in \mathbb{R}$ and $0 \leq a_t \leq 1$). The Importance Estimator is composed of a Convolutional Neural Network (CNN) and a bi-directional LSTM network (Bi-LSTM). The former is used to extract a set of feature vectors (one per frame) that represent the visual content of the frames ($\boldsymbol{X} = \{\boldsymbol{x_t}\}_{t=1}^T$). In particular, for feature extraction we use a model of GoogleNet [24] trained on ImageNet. The sequence of feature vectors is then processed by a bi-directional LSTM that models the temporal dependency over the sequence of frames in both forward and backward direction, and assigns a score to each video frame that represents its importance (frame-level scores $\boldsymbol{i} = \{i_t\}_{t=1}^T$ with $i_t \in \mathbb{R}$ and $0 \leq i_t \leq 1$). The computed scores about the aesthetic quality and importance of the frames' visual content are then fused based on an element-wise multiplication process (represented by the ⊗ symbol in Fig. 2), forming a new sequence of scores that capture information about both of the conducted assessments (frame-level scores $\boldsymbol{s} = \{s_t\}_{t=1}^T$ with $s_t \in \mathbb{R}$ and $0 \leq s_t \leq 1$). This sequence of scores is used by the Picking Mechanism of the Thumbnail Selector, that participates in a series of N training episodes as part of the applied reinforcement learning strategy. In each episode e, the Picking Mechanism selects a small set of frames as candidate thumbnails, based on a set of M discrete sampled actions over the group of video frames ($\boldsymbol{p_e} = \{p_k\}_{k=1}^M$ with $p_k \in \mathbb{N}$ and $1 \leq p_k \leq T$); each action indicates the selection or not of a video frame, and a frame can be selected more than once throughout the picking process. These actions lead to an updated sequence of frame-level scores per training episode, where the initially computed score for each selected frame is increased by 100% after a selection ($\boldsymbol{s'_e} = \{s'_t\}_{t=1}^T$).

The output of the thumbnail selection process (i.e., the set of selected candidate thumbnails $\boldsymbol{p_e} = \{p_k\}_{k=1}^M$ and the updated sequence of frame-level scores $\boldsymbol{s'_e} = \{s'_t\}_{t=1}^T$) is given as input to the Thumbnail Evaluator. The information about the selected frames is utilized by the Aesthetic Evaluator. Given this information, the latter computes an overall aesthetics score for the set of candidate thumbnails, as the average of the assigned aesthetics scores in the selected frames by the relevant mechanism of the Thumbnail Selector. This score forms the Aesthetics reward of the Thumbnail Evaluator for the current training episode ($r_{a_e} \in \mathbb{R}$). The updated sequence of frame-level scores, that also carries information about the choices of the Thumbnail Selector, is used to create a weighted version of the original feature vectors ($\boldsymbol{W_e} = \{\boldsymbol{w_t}\}_{t=1}^T$). This weighted version is given as input to the Representativeness Evaluator. The latter is composed of a Generator and a Discriminator that are being trained in an adversarial manner. The Generator is an LSTM-based Variational Auto-Encoder which tries to discover the underlying structure of the weighted data after the choices made by the Thumbnail Selector, and reconstruct the original data by drawing samples from a distribution over the learned latent space ($\boldsymbol{X'_e} = \{\boldsymbol{x'_t}\}_{t=1}^T$). The goal of this encoding-decoding process is to minimize the reconstruction error and produce a representation of the original video that fools the Discriminator. The latter is an LSTM that gets as input the original feature vectors $\boldsymbol{X}$ and the reconstructed feature

[1] Code retrieved from: https://github.com/bmezaris/fully_convolutional_networks

Proc. ACM ICMR 2021. Authors' accepted version. The final paper is avaiable at https://doi.org/10.1145/3460426.3463630

vectors in each training episode $\boldsymbol{X'_e}$, based on the Thumbnail Selector's choices and the following encoding-decoding process. Then, it defines a new latent representation for each of the aforementioned versions of feature vectors, and computes the reconstruction loss L_{rec} (scalar value) based on the proximity of these representations. The subtraction of the computed loss value from the unit ($1 - L_{rec}$) forms the Representativeness reward of the Thumbnail Evaluator for the current training episode ($r_{b_e} \in \mathbb{R}$). Finally, the computed rewards by the Aesthetics and Representativeness Evaluators are fused through an averaging operator (represented by the ⊘ symbol in Fig. 2). The output of this operation defines the overall reward for the current training episode ($r_e \in \mathbb{R}$). This reward represents the feedback of the Thumbnail Evaluator and is used to train the Thumbnail Selector through reinforcement learning.

At inference time the Thumbnail Evaluator gets as input the raw video content. Assuming a video of T frames, it produces a sequence of frame-level scores ($\boldsymbol{s'} = \{s'_t\}_{t=1}^{T}$) that signify each frame's suitability - according to the aesthetics and importance of its visual content - to be a video thumbnail. The analysis involves the processing of the video frames by the Aesthetics and Importance Estimators of the Thumbnail Selector. The former processes the frame sequence and assigns a score to every video frame according to the aesthetic quality of its visual content ($\boldsymbol{a} = \{a_t\}_{t=1}^{T}$). The latter represents the visual content of the video frames with the help of a pretrained CNN which extracts one feature vector per frame ($\boldsymbol{X} = \{\boldsymbol{x_t}\}_{t=1}^{T}$). The sequence of the extracted feature vectors is then processed by the trainable bi-directional LSTM which computes an importance score for every frame, based on its temporal dependency with the other frames of the video ($\boldsymbol{i} = \{i_t\}_{t=1}^{T}$). The calculated aesthetics and importance scores are combined via an element-wise multiplication procedure and form a sequence of fused scores ($\boldsymbol{s} = \{s_t\}_{t=1}^{T}$). Finally, based on the sequence of fused scores the Picking Mechanism of the Thumbnail Selector chooses a small set of frames based on a set of M discrete sampled actions over a multinomial distribution that follows the distribution of the fused data. These choices result in an updated sequence of frame-level scores with increased values for the selected video frames ($\boldsymbol{s'} = \{s'_t\}_{t=1}^{T}$). Finally, the top-scored frame or frames are selected as the video thumbnails.

3.3 Learning objectives and pipeline

The trainable parts of the developed architecture are indicated by the shaded boxes in Fig. 2. The learning objectives for training the Encoder, Decoder and Discriminator of the proposed architecture include: a prior loss (L_{prior}), a reconstruction loss (L_{rec}), the "original" (L_{ORIG}) and "summary" (L_{SUM}) losses, and the generator loss (L_{GEN}). For sake of space we provide a short explanation of these losses and refer the reader to [1, 18] for a more detailed description. Then, we present the applied episodic reinforcement learning approach for training the bi-directional LSTM component of the Importance Estimator of the architecture.

L_{prior} measures how much information is lost when using the Encoder's latent space to represent the prior distribution defined by the Variational Auto-Encoder that acts as the Generator of the Representativeness Evaluator. L_{rec} estimates the distance between the original and the reconstructed feature vectors, based on a learned latent representation in the last hidden layer of the Discriminator, that is part of the Representativeness Evaluator. L_{ORIG} and L_{SUM} relate to a label-based training approach (labels "1" and "0" denote the original and the reconstructed feature vectors for the adversarial part of our method) and are used to train the Discriminator; L_{ORIG} is used to minimize the difference between the computed probability and label "1" when the Discriminator gets the original feature vectors, and L_{SUM} is used to minimize the difference between the computed probability and label "0" when the Discriminator gets the thumbnail-based reconstructed feature vectors. Finally, L_{GEN} is used to minimize the difference between the probability computed by the Discriminator when the latter is fed with the reconstructed feature vectors and label "1", thus forcing the Generator to reconstruct a video that is hard to distinguish from the original.

Algorithm 1 The applied episodic REINFORCE algorithm for training the Importance Estimator of the developed architecture.

Notation: T is the number of video frames, M is the number of selected key-frames, N is the number of training episodes per epoch, a_p is the aesthetic score for frame p, b is a constant baseline that facilitates network's convergence, L is the loss, and MDIST is the multinomial distribution for action sampling given the set of scores $\boldsymbol{s} = \{s_t\}_{t=1}^{T}$; for the e^{th} episode: $\boldsymbol{p_e} = \{p_k\}_{k=1}^{M}$ is a vector with the indices of the selected key-frames, L_{rec_e} is the reconstruction loss, r_{a_e} is the aesthetics reward, r_{b_e} is the representativeness reward, r_e is the overall reward

Input: A training sample (video).

Output: The computed gradients for this training sample.

```
for $e = 1 \rightarrow N$ do
    # compute aesthetics reward: $r_{a_e} = \frac{1}{M} \sum_{k=1}^{M} a_{p_k}$
    # compute representativeness reward: $r_{b_e} = 1 - L_{rec_e}$
    # compute overall reward: $r_e = \frac{r_{a_e} + r_{b_e}}{2}$
    # compute logarithm of probability density function
    $log\_prob_e = MDIST.log\_prob(\boldsymbol{p_e})$
    # compute expected reward: $er_e = log\_prob_e\ (r_e - b)$
    # minimize negative expected reward: $L = -er_e$
# compute gradients: $L.backward()$
# update b based on moving average of received rewards
$b = 0.9b + 0.1\frac{1}{N}\sum_{e=1}^{N} r_e$
```

To train the Importance Estimator we apply an episodic REINFORCE algorithm, as implemented in [33] and described in Alg. 1. Given the e^{th} training episode, the computed aesthetics (r_{a_e}) and representativeness (r_{b_e}) rewards for the set of selected key-frames are combined (via an averaging operation) to form the overall reward for the episode (r_e). The latter is then used to compute the maximum expected reward based on the logarithm of the probability density function evaluated for the given sampled actions (selected key-frames) of the multinomial distribution, and a constant baseline b. The loss L that is used to train the Importance Estimator is formed so as to minimize the negative expected reward. After the end of the training episodes, the gradients are computed based on the accumulated loss value, and the baseline b is updated based on the moving average of the received rewards during the episodes. Based on this training strategy, the Importance Estimator learns a policy for scoring the video frames, by maximizing the expected rewards from the Thumbnail Evaluator.

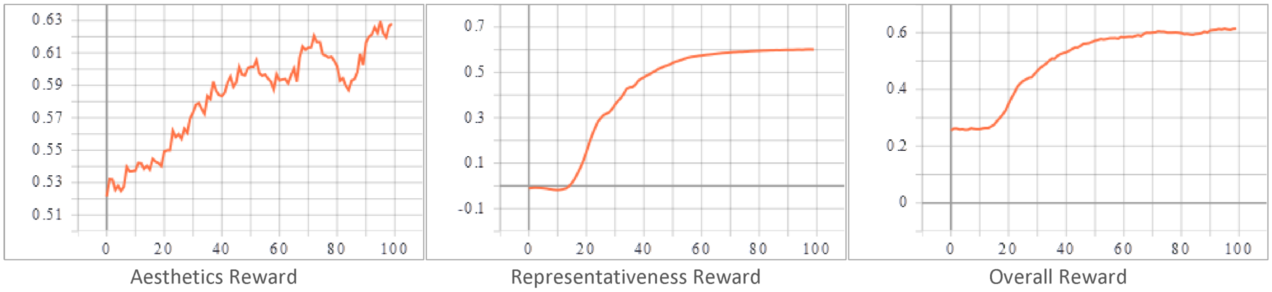

Figure 3: Reward curves for the proposed model. The horizontal axis in all plots indicates the epoch number. These curves show smooth training of the model, and the ability of the Thumbnail Selector to get higher rewards as the training proceeds.

With regards to the training pipeline, we follow a step-wise approach similar to the one in [1]. First we update the Encoder of the architecture. Then, having the Encoder updated, we proceed by updating the Decoder. Subsequently, having the aforementioned components updated (i.e., having some knowledge about the task), we update the Discriminator. Finally, based on the received feedback from the updated Discriminator, we update the Importance Estimator. The above described step-wise learning process allows all the different components to be trained effectively, and the Thumbnail Selector gets higher rewards as the training proceeds (see Fig. 3).

4 EXPERIMENTS

This section reports on the conducted experiments. It starts by presenting the utilized datasets and evaluation approach (Sec. 4.1), and continues by providing details about the implemented architecture and the applied training process (Sec. 4.2). Subsequently, it describes the findings of performance comparisons with other methods of the literature (Sec. 4.3). Finally, it discusses the setup and the outcomes of an ablation study that aims to assess the contribution of each adopted criterion for thumbnail selection (Sec. 4.4).

4.1 Datasets and evaluation approach

A study of the literature indicated the lack of a commonly-accepted protocol for evaluating video thumbnail selection. A few works perform assessments based on sets of proprietary or collected data and subjective human evaluations (e.g., [7, 17, 25, 32]). Other approaches rely on publicly-available data but differ in the way they estimate similarity among the selected and the ground-truth thumbnails for a given video. For example, [9] uses the OVP and Youtube datasets [8] and estimates similarity based on the Structural Similarity Index (SSIM) and a predefined threshold. [23] uses the Yahoo dataset and estimates similarity using the SIFTFlow algorithm [15] and an experimentally-defined threshold. A recent multimodal approach [28] uses a subset of the Yahoo dataset, computes the Mean Squared Error among the extracted representations in a latent space, and reports results for different values of the computed distance.

Given the above, we choose to assess the performance of the proposed method using the datasets and evaluation protocol adopted in [9]. So, in terms of data we utilize the OVP and Youtube datasets. Each of these datasets is made of 50 videos with diverse video content, such as documentaries, historical, and lecture videos (OVP dataset) and news, TV-shows and home videos (Youtube dataset). The video duration ranges from 46 sec. to 3.5 min. in the case of the OVP videos, and from 9 sec. to approximately 11 min. in the case of Youtube videos. Each video of these datasets has been annotated by 5 users, where each user was asked to select a set of representative key-frames. Following the evaluation approach in [9] that relates to the thumbnail selection task, we consider the top-3 selected key-frames among all annotators for a given video as the ground-truth thumbnails for this video. As a side note, through this procedure some videos are associated with more than 3 ground-truth thumbnails, due to the existence of more than 3 key-frames with the same ranking according to the number of selections made by the human annotators. Finally, in terms of the utilized measure we quantify the performance of the proposed method based on a top-3 matching process - i.e., the top-3 selected thumbnails by our method against the top-3 ground-truth thumbnails - similarly to [9]. In addition, we measure the performance when considering only the top-1 machine- and user-selected thumbnails for each video.

4.2 Implementation details

All videos were downsampled to 2 fps. The aesthetics quality of each video frame was computed as the softmax of the values in the final layer of the utilized FCN architecture of [3]. To represent the visual content of the video frames, we used the output of pool5 layer of GoogleNet [24] trained on ImageNet, and extracted one feature vector (containing 1024 values) per frame. The trainable part of the Importance Estimator is made of a 2-layer bi-directional LSTM with 512 hidden units. All the different parts of the Representativeness Evaluator are 2-layer LSTMs with 512 hidden units. Training is performed in a full-batch mode using the Adam optimizer. The number of candidate thumbnails M is set equal to 10, and the same holds for the number of episodes N per training epoch, $N = 10$. The learning rate for all components but the Discriminator is 10^{-4} and for the latter one is 10^{-5}. Training stops after a maximum number of epochs (100 in our case). As a well-trained model we select the one that maximizes the overall reward on the entire set of training data (see the rightmost graph of Fig. 3). With respect to the used data, we adopted the typical learning setting in most SoA video summarization works (e.g., [1, 18, 31]) where the used dataset is split into two non-overlapping sets; a training set containing 80% of data, and a testing set made of the remaining 20% of data. In the case of Youtube, we excluded 10 cartoon videos, since the utilized networks for feature extraction (the GoogleNet of [24]) and aesthetic quality estimation (the FCN architecture of [3]) cannot provide meaningful representations and aesthetics measurements for the content of these videos. Finally, driven by the recent reportings in [2] about the varying difficulty of the different randomly-created splits of data of other relevant datasets (that are extensively used for evaluating video summarization algorithms), to reduce the impact of the utilized data split for training and testing our method we run our experiments on 10 different randomly-created splits and in the following we report the average performance over these runs.

4.3 Performance comparisons

The proposed approach is compared against a baseline that selects video thumbnails randomly, and a set of SoA approaches for video thumbnail selection and summarization from the literature. To estimate the performance of the baseline, we randomly scored the video frames of each test video based on a uniform distribution of probabilities. Then, we evaluated the performance of this baseline on a given test video, by comparing the top-1 and top-3 scoring frames with the defined ground-truth thumbnails for this video. This experiment was repeated 100 times for the videos of each utilized test set in our experiments, and the overall average score over these iterations and over the 10 different data splits is reported.

Table 1 presents the experimental outcomes when using the top-3 selected key-frames by all human annotators, as the ground-truth thumbnails for each video. The reported values express "Precision at k" in percentages. When $k = 3$, we implement the evaluation protocol of [9], that compares the top-3 automatically-selected and the top-3 ground-truth thumbnails. The reported results in this table show that the proposed approach is the best-performing one in both datasets and both experimental settings. More specifically, when using the top-3 selected thumbnails for evaluation our method outperforms all the other reported approaches in both OVP and Youtube datasets. Nevertheless, we should stress that the reported values for the first three compared methods (presented in [9, 18, 23]) are the ones reported in [9]; their experimental reproduction as part of this work was not feasible due to the limited implementation details provided in [9]. For example there are no details about the used CNN for feature extraction, the split of data into training and testing samples, and the number of iterations (if any) of the conducted experiments. In the most challenging scenario where only one thumbnail is selected and compared with the 3-thumbnails ground-truth (P@1), the proposed method is by far more competitive than the considered baseline, showing a performance increase by approximately 100% compared to random selection. In addition our method seems to be more effectively-tailored to the video thumbnail selection task compared to the SoA video summarization algorithm from [1], that was evaluated under the same experimental conditions using its publicly-available implementation[2].

Table 2 reports our findings when only the top-1 selected key-frame by all human annotators is used as the ground-truth thumbnail for each video. Our comparisons involve the baseline (random selection), and the SoA video summarization method from [1]. Even in this more demanding scenario - which is much closer to the users' needs when mature video thumbnail selection technologies will be used in practice - our method performs significantly better than the baseline (being two times more precise when selecting a single thumbnail) and clearly exceeds the performance of a SoA video summarization method in both datasets and evaluation settings.

Table 1: Performance comparison when using the top-3 selected key-frames by the human annotators as the ground-truth thumbnails for each video. P@k denotes "Precision at k" (as percentage). Best scores in bold font.

	OVP		Youtube	
	P@1	P@3	P@1	P@3
Baseline (random)	15.79	32.51	7.53	17.94
Mahasseni et al. [18]	-	7.80	-	11.34
Song et al. [23]	-	11.72	-	16.47
Gu et al. [9]	-	12.18	-	18.25
Apostolidis et al. [1]	15.00	24.00	8.75	15.00
Proposed approach	**31.00**	**40.00**	**15.00**	**20.00**

Table 2: Performance comparison when using the top-1 selected key-frames by the human annotators as the ground-truth thumbnails for each video. P@k denotes "Precision at k" (as percentage). Best scores in bold font.

	OVP		Youtube	
	P@1	P@3	P@1	P@3
Baseline (random)	6.36	16.66	4.23	9.98
Apostolidis et al. [1]	7.00	14.00	6.25	8.75
Proposed approach	**17.00**	**21.00**	**10.00**	**16.25**

Table 3: The variants of the proposed approach, that were examined in the ablation study.

	Aesthetic quality estimations		Representativ. estimations
	In frame selection	As a reward	
Variant #1	✓	✓	X
Variant #2	X	X	✓
Variant #3	✓	X	✓
Variant #4	X	✓	✓
Proposed approach	✓	✓	✓

With regards to space requirements, the memory footprint of the network is 1.2GB. Concerning time requirements, on a PC with an i7-3770K CPU, 32GB RAM and an RTX2080Ti GPU, the time needed for training using the OVP and Youtube datasets is 1.5 and 2.4 min. per epoch respectively (average values over the 10 used data splits). At inference time, thumbnail selection takes less than 0.3 sec. per video. Finally, feature extraction and aesthetics scoring estimation takes about 6 and 30 msec. per frame, respectively.

4.4 Ablation study

To assess the impact of each of the adopted criteria for thumbnail selection, we conduct an ablation study which includes the following variants of the proposed approach (also presented in Table 3):

[2] https://github.com/e-apostolidis/AC-SUM-GAN

Table 4: Ablation study based on the performance (P@k (%) with k = 1, 3) of four variants of the proposed approach, on the OVP and Youtube datasets. Best scores in bold font, second best scores underlined.

	Thumbnail selection criteria			OVP				Youtube			
	Aesthetics estimations		Represent. estimations	Using top-3 human selections		Using top-1 human selections		Using top-3 human selections		Using top-1 human selections	
	Frame picking	Reward		P@1	P@3	P@1	P@3	P@1	P@3	P@1	P@3
Baseline (random)	-	-	-	15.79	32.51	6.36	16.66	7.53	17.94	4.23	9.98
Variant #1	✓	✓	X	16.00	20.00	8.00	12.00	6.00	17.50	5.00	7.50
Variant #2	X	X	✓	20.00	30.00	8.00	13.00	10.00	18.75	3.75	8.75
Variant #3	✓	X	✓	12.00	36.00	3.00	18.00	10.00	18.75	6.25	12.50
Variant #4	X	✓	✓	30.00	39.00	**18.00**	**23.00**	13.75	16.25	**10.00**	12.50
Proposed approach	✓	✓	✓	**31.00**	**40.00**	17.00	21.00	**15.00**	**20.00**	**10.00**	**16.25**

- Variant #1 does not measure the representativeness of the selected set of candidate thumbnails. Thumbnail selection is based only on the computed scores about the aesthetics of the visual content of video frames ($\boldsymbol{a} = \{a_t\}_{t=1}^{T}$ in Fig. 2). The received reward is maximized by simply selecting the top-k scored frames (in our experiments k equals to 1 and 3).
- Variant #2 does not make any estimates about the aesthetics of the visual content of video frames. The Aesthetics Estimator and the Aesthetics Evaluator of the proposed architecture (see Fig. 2) are completely missing. Thumbnail selection relies solely on measurements about the representativeness of the set of candidate thumbnails.
- Variant #3 uses the computed scores about the aesthetics of the visual content of video frames ($\boldsymbol{a} = \{a_t\}_{t=1}^{T}$ in Fig. 2) only for computing the set of scores that capture information about both aesthetics and importance ($\boldsymbol{s} = \{s_t\}_{t=1}^{T}$), that subsequently affect the frame selection process. No information about the aesthetics is utilized by the Thumbnail Evaluator, and the overall reward after a training episode (r_e) equals to the computed representativeness reward (r_{b_e}).
- Variant #4 uses the computed scores about the aesthetics of the visual content of video frames ($\boldsymbol{a} = \{a_t\}_{t=1}^{T}$ in Fig. 2) only to estimate the overall aesthetics score of the selected set of candidate thumbnails, and uses this score as a reward (see the Aesthetics Evaluator component in Fig. 2). No information about the aesthetics is utilized by the Frame Picking Mechanism, and frame selection is affected only by the computed scores about the visual importance of the video frames ($\boldsymbol{s} = \{s_t\}_{t=1}^{T}$ equals to $\boldsymbol{i} = \{i_t\}_{t=1}^{T}$).

Based on the results reported in Table 4 we make the following observations: The developed method is the best performing one in most considered settings (6 out of 8 in total), and the second best in the remaining ones (by a very small margin from the best). Extracting and using information about the aesthetics of the visual content only as part of the received reward signal, also allows the Thumbnail Selector to gain good knowledge about the task. The corresponding variant (Variant #4) is the second best performing algorithm in most experimental settings. When the aesthetic quality is not taken under consideration for rewarding the Thumbnail Selector (Variant #3) or it is completely ignored (Variant #2), the performance deteriorates in most cases. Finally, when no estimates are being made with regards to the representativeness of the visual content and thumbnail selection relies solely on the aesthetics (Variant #1), the performance is comparable with the performance of random selection. The above show that measuring both representativeness and aesthetic quality of the visual content and combining this knowledge as proposed, leads to the best performance.

5 CONCLUSIONS

In this work we proposed a new approach for video thumbnail selection. This approach is based on a deep-learning network architecture and a training strategy that combines adversarial and reinforcement learning. The selection is based on assessments with regards to the representativeness and aesthetic quality of the visual content of the video frames. The former is estimated with the help of an adversarially-trained discriminator and the latter is computed using a pretrained Fully Convolutional Network. An overall score is formed based on the outcome of these assessments and used as a reward to train the thumbnail selector based on the principles of reinforcement learning. Experiments on two benchmark datasets (OVP and Youtube) showed the advanced performance of the proposed approach against other SoA video thumbnail selection or summarization methods. Finally, an ablation study signified the importance of aesthetics for the video thumbnail selection task, and documented the effectiveness of the proposed approach for selecting representative and aesthetically-pleasing video thumbnails.

ACKNOWLEDGMENTS

This work was supported by the EU Horizon 2020 programme under grant agreement H2020-832921 MIRROR, and by EPSRC under grant No. EP/R026424/1.

REFERENCES

[1] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras. 2020. AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology* (2020), 1–1. https://doi.org/10.1109/TCSVT.2020.3037883

[2] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. 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)*. Association for Computing Machinery, New York, NY, USA, 1056–1064. https://doi.org/10.1145/3394171.3413632

[3] K. Apostolidis and V. Mezaris. 2019. Image Aesthetics Assessment Using Fully Convolutional Neural Networks. In *Proc. of 25th Int. Conf. on MultiMedia Modeling (MMM 2019)*. Springer International Publishing, Cham, 361–373.

[4] N. Arthurs and S. Birnbaum. 2017. *Selecting Youtube Video Thumbnails via Convolutional Neural Networks*. Technical Report. Stanford.

[5] L. Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[6] J. F. Brenner, B. S. Dew, J. B. Horton, T. King, P. W. Neurath, and W. D. Selles. 1976. An automated microscope for cytologic research a preliminary evaluation. *Journal of Histochemistry & Cytochemistry* 24, 1 (1976), 100–111.

[7] J. Choi and C. Kim. 2016. A Framework for Automatic Static and Dynamic Video Thumbnail Extraction. *Multimedia Tools and Applications* 75, 23 (Dec. 2016), 15975–15991. https://doi.org/10.1007/s11042-015-2909-6

[8] S. E. F. de Avila, A. P. B. Lopes, A. da Luz Jr., and A. de A. Araújo. 2011. VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method. *Pattern Recognition Letters* 32, 1 (Jan. 2011), 56–68.

[9] H. Gu and V. Swaminathan. 2018. From Thumbnails to Summaries-A Single Deep Neural Network to Rule Them All. In *Proc. of the 2018 IEEE Int. Conf. on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME.2018.8486533

[10] J. Harel, C. Koch, and P. Perona. 2006. Graph-Based Visual Saliency. In *Proc. of the 19th Int. Conf. on Neural Information Processing Systems (NIPS'06)*. MIT Press, Cambridge, MA, USA, 545–552.

[11] X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan. 2019. Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks. In *Proc. of the 27th ACM Int. Conf. on Multimedia (MM '19)*. ACM, New York, NY, USA, 2296–2304. https://doi.org/10.1145/3343031.3351056

[12] Y. Jung, D. Cho, D. Kim, S. Woo, and I.-S. Kweon. 2019. Discriminative Feature Learning for Unsupervised Video Summarization. In *Proc. of the 2019 AAAI Conf. on Artificial Intelligence (AAAI 2019)*.

[13] H. Lian, Xiao-Qiang Li, and Bo Song. 2011. Automatic video thumbnail selection. In *Proc. of the 2011 Int. Conf. on Multimedia Technology*. 242–245. https://doi.org/10.1109/ICMT.2011.6002001

[14] C. Liu, Q. Huang, and S. Jiang. 2011. Query sensitive dynamic web video thumbnail generation. In *Proc. of the 2011 18th IEEE Int. Conf. on Image Processing (ICIP)*. 2449–2452. https://doi.org/10.1109/ICIP.2011.6116155

[15] C. Liu, J. Yuen, and A. Torralba. 2011. SIFT Flow: Dense Correspondence across Scenes and Its Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 5 (2011), 978–994. https://doi.org/10.1109/TPAMI.2010.147

[16] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma. 2007. Dual Cross-Media Relevance Model for Image Annotation. In *Proc. of the 15th ACM Int. Conf. on Multimedia (MM' 07) (MM '07)*. Association for Computing Machinery, New York, NY, USA, 605–614. https://doi.org/10.1145/1291233.1291380

[17] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 3707–3715. https://doi.org/10.1109/CVPR.2015.7298994

[18] B. Mahasseni, M. Lam, and S. Todorovic. 2017. Unsupervised Video Summarization with Adversarial LSTM Networks. In *Proc. of the 2017 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2982–2991.

[19] N. Murray, L. Marchesotti, and F. Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2408–2415. https://doi.org/10.1109/CVPR.2012.6247954

[20] K. Pretorious and N. Pillay. 2020. A Comparative Study of Classifiers for Thumbnail Selection. In *Proc. of the 2020 Int. Joint Conf. on Neural Networks (IJCNN)*. 1–7. https://doi.org/10.1109/IJCNN48605.2020.9206951

[21] J. Ren, X. Shen, Z. Lin, and R. Měch. 2020. Best Frame Selection in a Short Video. In *Proc. of the 2020 IEEE Winter Conf. on Applications of Computer Vision (WACV)*. 3201–3210. https://doi.org/10.1109/WACV45572.2020.9093615

[22] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. of the 3rd Int. Conf. on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*, Y. Bengio and Y. LeCun (Eds.). http://arxiv.org/abs/1409.1556

[23] Y. Song, M. Redi, J. Vallmitjana, and A. Jaimes. 2016. To Click or Not To Click: Automatic Selection of Beautiful Thumbnails from Videos. In *Proc. of the 25th ACM Int. on Conf. on Information and Knowledge Management (CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 659–668. https://doi.org/10.1145/2983323.2983349

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *Proc. of the 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[25] C. Tsao, J. Lou, and H. H. Chen. 2019. Thumbnail Image Selection for VOD Services. In *Proc. of the 2019 IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR)*. 54–59. https://doi.org/10.1109/MIPR.2019.00018

[26] A. B. Vasudevan, M. Gygli, A. Volokitin, and L. Van Gool. 2017. Query-Adaptive Video Summarization via Quality-Aware Relevance Estimation. In *Proc. of the 25th ACM Int. Conf. on Multimedia (MM' 17) (MM '17)*. Association for Computing Machinery, New York, NY, USA, 582–590. https://doi.org/10.1145/3123266.3123297

[27] Y. Gao, T. Zhang, and J. Xiao. 2009. Thematic video thumbnail selection. In *Proc. of the 2009 16th IEEE Int. Conf. on Image Processing (ICIP)*. 4333–4336. https://doi.org/10.1109/ICIP.2009.5419128

[28] Z. Yu and N. Shi. 2020. A Multi-modal Deep Learning Model for Video Thumbnail Selection. arXiv:2101.00073 [cs.CV]

[29] L. Yuan, F. E. H. Tay, P. Li, and J. Feng. 2020. Unsupervised Video Summarization With Cycle-Consistent Adversarial LSTM Networks. *IEEE Transactions on Multimedia* 22, 10 (2020), 2711–2722. https://doi.org/10.1109/TMM.2019.2959451

[30] Y. Yuan, L. Ma, and W. Zhu. 2019. Sentence Specified Dynamic Video Thumbnail Generation. In *Proc. of the 27th ACM Int. Conf. on Multimedia (MM '19 (MM '19)*. Association for Computing Machinery, New York, NY, USA, 2332–2340. https://doi.org/10.1145/3343031.3350985

[31] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. 2016. Video Summarization with Long Short-Term Memory. In *Proc. of the European Conf. on Computer Vision 2016 (ECCV)*, B. Leibe, J. Matas, N. Sebe, and M. Welling (Eds.). Springer International Publishing, Cham, 766–782.

[32] W. Zhang, C. Liu, Q. Huang, S. Jiang, and W. Gao. 2012. A Novel Framework for Web Video Thumbnail Generation. In *Proc. of the 2012 Eighth Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*. 343–346. https://doi.org/10.1109/IIH-MSP.2012.89

[33] K. Zhou and Y. Qiao. 2018. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. In *Proc. of the 2018 AAAI Conf. on Artificial Intelligence (AAAI 2018)*.