Multiple Pairwise Ranking Networks for Personalized Video Summarization

Yassir Saquil[1] Da Chen*[2] Yuan He[2] Chuan Li[3] Yong-Liang Yang[1]

[1]University of Bath [2]Alibaba Group [3]Lambda Labs

Abstract

In this paper, we investigate video summarization in the supervised setting. Since video summarization is subjective to the preference of the end-user, the design of a unique model is limited. In this work, we propose a model that provides personalized video summaries by conditioning the summarization process with predefined categorical user labels referred to as preferences. The underlying method is based on multiple pairwise rankers (called Multi-ranker), where the rankers are trained jointly to provide local summaries as well as a global summarization of a given video. In order to demonstrate the relevance and applications of our method in contrast with a classical global summarizer, we conduct experiments on multiple benchmark datasets, notably through a user study and comparisons with the state-of-art methods in the global video summarization task.

1. Introduction

Video summarization is an important subfield of video understanding. It aims to provide the end-user with a synopsis of the original video capturing only the relevant content. Various applications can benefit from video summarization, including semantic video editing and content filtering in particular. Moreover, the summary can also be used as a preprocessing step by excluding the unnecessary content and thus reducing the length and processing time of the video for downstream tasks such as action recognition.

Video summarization is often intertwined with highlight detection, which can be formulated as a subset selection problem based on a learnt model that assigns an importance score for each video frame or segment. In contrast, video summarization seeks a synopsis that not only contains the video highlight, but also satisfies other criteria, such as the diversity, representativeness, visual and semantic coherence of the summary. In addition, the ability for storytelling and adaptability to the context is often considered. A substantial work has been done in different topics in this research area notably in highlight detection and video summarization. Many of these works are based on a novel specific key insight or heuristic that takes into consideration one or more of summarization criteria [38, 13, 12, 73, 57, 51] along with a particular model setting such as supervised [15, 67, 11, 39], weakly supervised [19, 25, 37, 42] and unsupervised [62, 33, 73, 23] learning.

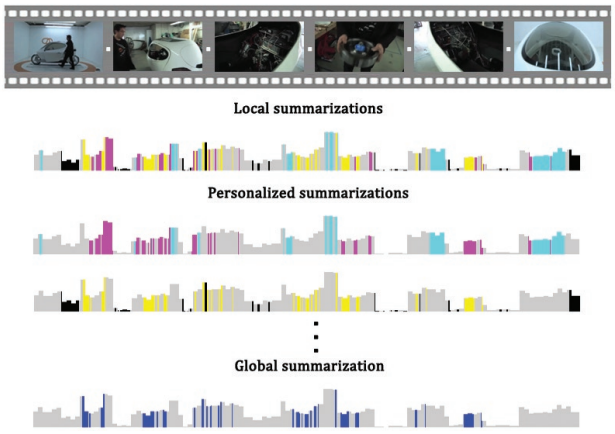

Figure 1. Given the test video 9 from TVSum dataset and its frame-level importance scores (gray), our Multi-ranker method trained with 4 preferences (cyan, magenta, yellow, black) can generate local/personalized summaries according to a subset of preferences and a global summary akin to classical summarization methods.

In practice, what is meaningful in a video is a subjective matter that depends on the end user perspective, which represents one of the main challenges in this research topic [48, 53]. However, few works in the literature were interested in exploring the possibility of customizing the generated summary [10, 41, 46, 54]. Since there is no consensus on what constitutes a global summary, suggesting a model that generates a unique summary is restrictive for general users due to the diversity of their perspectives and opinions. Instead, designing a model that can provide a set of summaries for the users to select is likely to satisfy specific user preferences. For instance, in a basketball match video containing many play actions (shooting, dribbling, slam dunk,

*corresponding author

layup, etc.), a unique global summary might include all or some of these actions. However, each user has particular action preference, and we believe considering such preference will enable a more flexible personalized summarization.

For these purposes, we propose in this work a novel ranking based video summarization model, that is constituted of multiple sub-ranking models that are trained using pairwise comparisons between the importance scores of video segments to provide local summaries with respect to each predefined preference by ranking important segments higher than unimportant ones. Moreover, the sub-ranking models are jointly trained so that the maximum of their predicted ranking scores ensures a unique global summary akin to a standard ranking model trained using pairwise comparisons. As a consequence, our model is capable of predicting a global ranking score and a set of local ranking scores according to each preference for a given video segment, enabling the possibility of interacting with the model by selecting one preference for a local summary generation, some preferences for a personalized summary generation, or all preferences for a global summary generation as illustrated in Figure 1. We demonstrate the relevance and applications of our method through quantitative and qualitative experiments in benchmark datasets.

In summary, our contributions are two-fold: 1) We show that a pairwise ranking based model can achieve state-of-art results in the task of supervised video summarization. 2) We propose a multiple pairwise ranking model endowed with a training scheme for generating a global summary as well as local and personalized summaries with respect to predefined preferences that the end-user can interact with.

2. Related Work

Video Highlight Detection In early sports video highlight detection works, different models are presented based on the audio features classification [44], visual features classification [52] or both of them [59, 55].

Most recent works formulate it as a pairwise ranking problem in different training settings. In the supervised setting, [51] proposed a pairwise ranker with latent variables that accounts for the noise and variation in the data while assuming that domain specific edited videos are more likely to contain highlights. [63] proposed a deep pairwise ranking model based on a two-stream network structure with video timelapse and skimming applications. [21] further integrated an attention model for better highlight predictions. [15] focused on generating animated GIFs using a ranking model with an adaptive Huber loss. Instead of standard videos, [64] tackled highlight detection from 360° videos by suggesting a suitable ranking model.

Fewer works are presented in other settings, [19] focused on the weakly supervised setting where only video event labels are provided. In contrast, [62] suggested an unsupervised approach using edited videos only. Based on the insight that short videos are more focused on highlight than long ones, [57] introduced a pairwise ranker that scores segments of short videos higher than longer videos.

Video Summarization While highlight detection focuses on finding relevant content in videos, video summarization imposes more constraints on the form of generated synopsis for application purposes. For egocentric video summarization, [60] proposed a model using gaze tracking information, while [29] suggested an approach driven by predicted important people and objects. [58, 27, 32] focused on the generation of a storyline representation of the summary.

For general videos, few works introduced non-learning methods. [34] proposed a motion based model using optical flow. [35] utilized a graph representation focusing on the content coverage and visual quality. [72] proposed an online method for quasi real-time summarization. [38] made a collaborative summary using knowledge extracted from similar videos. [8] defined the video shot importance by its visual co-occurrence in other videos.

Learning-based video summarization methods can be categorized according to their training settings. In the supervised setting, the classical methods first segment the video then estimate visual interestingness per segment using a set of features [14], mixtures of objectives [13], or a trained classifier [43, 40]. Other works model the interdependencies between frames using recurrent neural networks (RNNs) for better summarization. [67] proposed a LSTM model, while [70, 71] proposed a hierarchical RNN that perform both segmentation and summarization. To mitigate the computational cost, [11] proposed a self-attention network. Additional formulations have also been introduced such as sequential subset selection [12], graph modelling [39], and subset structure transfer [66].

In the weakly supervised setting, most of the approaches use the web prior information to enhance summarization task [6, 37, 49, 25, 26, 30, 42], while in unsupervised learning setting, many proposed methods learn from preexisting video cues [33, 65, 73, 22, 18, 68, 23]. We refer the reader to [3] for an in-depth discussion on these settings' methods.

Personalized Content Summarization Many works have investigated the possibility of customizing video summarization. Early works relied on meta-data and user profile. [20] trained an importance classifier by extracting features from the meta-data according to a specific user. [1] proposed to map between a user profile and multimedia features. [4] built users profiles and associated them with presentation media (e.g. video metadata, images) for personalized summarization of sports video.

Recent works focused more on using textual queries. [61] proposed to generate short summaries based on the user interaction using natural language questions with implicit constraints. [31] proposed a multi-task embedding network

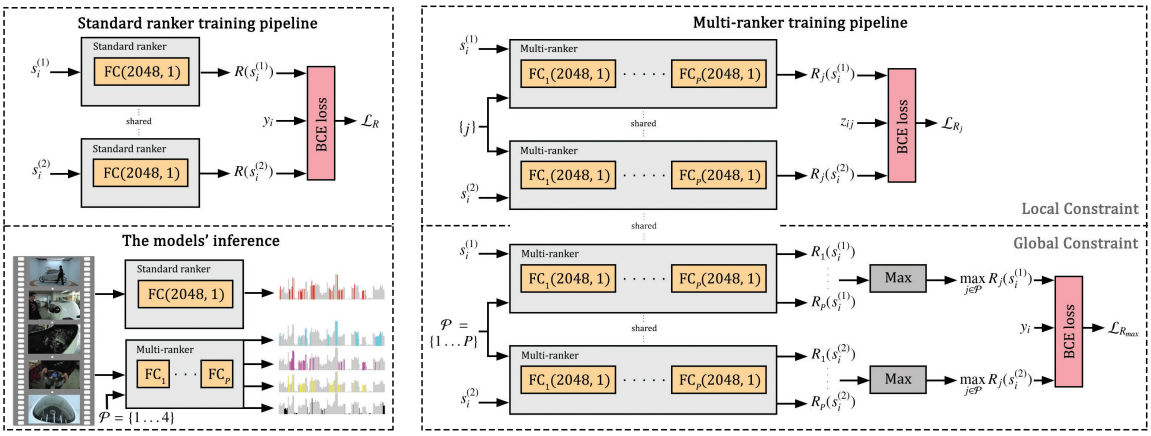

Figure 2. Standard ranker and Multi-ranker training pipelines and inference such that $FC(2048,1)$, the fully-connected layer with 2048 input features dimension and BCE, the binary cross-entropy loss. The grey color represents video frame-level GT importance scores and the top 15% predicted ranking scores are colored in the histogram to represent the summaries.

which incorporates information in form of title, description, query with visual content for semantic selection. [47] suggested a memory network that attends the user query onto different video frames. [69] proposed a query conditioned generative model where the generator learns a joint representation of user query and video content. [46, 54, 56] proposed methods to generate diverse, representative, and relevant summaries to the input text query.

In another spectrum, [10, 41] introduced personalization methods based on the user history. [10] suggested a ranking model conditioned on the user history represented by the previously selected highlights of the user. [41] trained highlight detection and history encoder networks that are interacted to provide frame-level highlight scores given the user input and previous highlights history. [9] proposed an active summarization method that interactively gathers user preference as feedback while creating the summary.

Our work can be considered as a personalized summarization using predetermined preferences. These preferences are segment-based labelling denoting a semantic meaning defined by the data. At test time, the user can interact with the trained model by selecting the preferences to include in the custom summary.

3. Approach

In this section, we first outline the pairwise ranking model for global supervised summarization, then we show how to incorporate it in the design of a multiple pairwise ranking model for local and personalized summarizations.

3.1. Pairwise Ranking Model for Global Summarization

Ranking-based models for video summarization are well studied in the literature [57, 15, 63] where the general aim is to learn a ranking function R that associates high ranking scores to important video segments and to build a summary by selecting the top-ranked segments. In this work, we formulate the summarization task as a classification problem and trained a pairwise ranker R using CNN features as proposed in [5, 50]. The method consists of learning to classify a pair of video segment features, $s^{(1)}$ and $s^{(2)}$ according to their GT (ground truth) importance ordering, $s^{(1)} \prec s^{(2)}$, $s^{(1)} \sim s^{(2)}$ or $s^{(1)} \succ s^{(2)}$.

Formally, given a video composed by n segments with features and their GT importance scores as $S = \{(s_1, l_1), \ldots, (s_n, l_n)\}$, we construct a dataset with pairwise comparisons $\{(s_i^{(1)}, s_i^{(2)}, y_i)\}_{i=1}^{N}$ of size N, and define the ranking loss as follows:

$$\mathcal{L}_R(s_i^{(1)},\ s_i^{(2)},\ y_i) = -y_i \log[\sigma(R(s_i^{(1)}) - R(s_i^{(2)}))] \\ -(1-y_i)\log[1-\sigma(R(s_i^{(1)}) - R(s_i^{(2)}))]), \quad (1)$$

such that $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, $(s_i^{(1)}, s_i^{(2)})$ is pair of segment features, and $y_i \in \{0, 0.5, 1\}$ is the comparison label that denotes the importance ordering and defined as: $y_i = [l_i^{(1)} > l_i^{(2)}] + \frac{1}{2}[l_i^{(1)} = l_i^{(2)}]$ with [.] is Iverson bracket and $(s_i^{(1)}, l_i^{(1)}), (s_i^{(2)}, l_i^{(2)}) \in S$.

3.2. Multiple Pairwise Ranking Model for Personalized Summarization

By training a pairwise ranking model R, we are capable of generating a unique summary. However, due to the subjective nature of summarization task, the applications of a global model are limited and a personalized model is much desirable thanks to its range of options, such as providing summary according to an additional cue and interactivity with the end-user. For these aims, a conditional model is an appropriate choice, for instance, [10, 41] proposed a conditional model on the user history to provide a custom sum-

mary. In contrast, we opt in this work for a model using supplemental categorical labelling that we aliased as preferences. The categorical label or preferences can represent many cues depending on the target dataset and available labelling, such as the action recognition labels or simply k-means clustering predicted labels.

Given a set of preferences $\mathcal{P} = \{1 \cdots P\}$, we introduce a multiple ranking model titled, Multi-ranker, which consists of a set of sub-rankers $\{R_j\}_{j=1}^{P}$ that are jointly trained so the local summaries conform with the preferences and global summary aggregates the sub-rankers scores. Formally, given a video as a set of n segments' features with their GT importance scores and preferences $S = \{(s_1, l_1, p_1), \ldots, (s_n, l_n, p_n)\}$, we construct a pairwise comparisons dataset $\{(s_i^{(1)}, s_i^{(2)}, y_i, z_{ij})\}_{(i,j)}$ with $(i, j) \in \{1 \ldots N\} \times \{1 \ldots P\}$, and define the local ranking loss $\mathcal{L}_{R_j}$ with respect to preference j associated with sub-ranker R_j as follows:

$$\mathcal{L}_{R_j}(s_i^{(1)},\ s_i^{(2)},\ z_{ij}) = -z_{ij} \log[\sigma(R_j(s_i^{(1)}) - R_j(s_i^{(2)}))] \\ -(1 - z_{ij}) \log[1 - \sigma(R_j(s_i^{(1)}) - R_j(s_i^{(2)}))]), \quad (2)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, $(s_i^{(1)}, s_i^{(2)})$ is a pair of segment features and $z_{ij} \in \{0, 0.5, 1\}$ is the local comparison label that denotes the importance ordering with respect to preference j and ensures that a segment s_i is of high importance only if l_i is high and $p_i = j$ which can be formulated as follows:

$$z_{ij} = [([p_i^{(1)} = j]\ l_i^{(1)}) > ([p_i^{(2)} = j]\ l_i^{(2)})] \\ +\frac{1}{2}[([p_i^{(1)} = j]\ l_i^{(1)}) = ([p_i^{(2)} = j]\ l_i^{(2)})]. \quad (3)$$

Additionally, in order to ensure that the max-aggregation of sub-rankers $\{R_j\}$ scores is equivalent to a global pairwise ranker R, we define the global ranking loss $\mathcal{L}_{R_{max}}$ similarly to $\mathcal{L}_R$ as follows:

$$\mathcal{L}_{R_{max}}(s_i^{(1)}, s_i^{(2)}, y_i) = -y_i \log[\sigma(\max_{j \in \mathcal{P}} R_j(s_i^{(1)}) - \max_{j \in \mathcal{P}} R_j(s_i^{(2)}))] \\ -(1 - y_i) \log[1 - \sigma(\max_{j \in \mathcal{P}} R_j(s_i^{(1)}) - \max_{j \in \mathcal{P}} R_j(s_i^{(2)}))]). \quad (4)$$

Finally, putting the global and local losses together we obtain the Multi-ranker loss $\mathcal{L}_{R_{multi}}$ as follows:

$$\mathcal{L}_{R_{multi}}(s_i^{(1)},\ s_i^{(2)},\ y_i,\ z_{ij}) = \lambda\ \mathcal{L}_{R_{max}}(s_i^{(1)},\ s_i^{(2)},\ y_i) \\ +(1 - \lambda)\ \mathcal{L}_{R_j}(s_i^{(1)},\ s_i^{(2)},\ z_{ij}), \quad (5)$$

where $(s_i^{(1)}, s_i^{(2)})$ is a pair of segment features, $z_{ij}, y_i \in \{0, 0.5, 1\}$ are the local and global comparison labels and λ is a hyperparameter that balances between local and global summarization of the Multi-ranker model.

By leveraging the Multi-ranker model and its training scheme illustrated in Figure 2, we can perform the following three main tasks:

- Global summarization similar to state-of-the-art supervised summarization methods, where the global predicted ranking score for segment s_i is $\max_{j \in \mathcal{P}} R_j(s_i)$
- Local summarization with respect to a specific preference j, where the local predicted ranking score for segment s_i is $R_j(s_i)$
- Personalized summarization with respect to a specific subset of preferences $\mathcal{P}_s \in 2^{\mathcal{P}} \backslash \{\}$, where the custom predicted ranking score for segment s_i is $\max_{j \in \mathcal{P}_s} R_j(s_i)$

We note that the possibility of separately training a global ranker R and local sub-rankers $\{R_j\}$ to perform local and global summarization tasks, shows that the strength of our Multi-ranker model lies in its training scheme that correlates between local sub-rankers to create a personalized and eventually a global summary. In such a scenario, the independent models provide $|\mathcal{P}|+1$ summaries, while Multi-ranker model generates $2^{|\mathcal{P}|}-1$ summaries thanks to the possibility of selecting different preference combinations.

Lastly, different from [10, 41] that require the user history, at test time, Multi-ranker only requires the input video and preference selection from the user thanks to the preference modelling we opt for. However, our method requires as many sub-rankers as preferences. We believe this is not a computational issue, since each ranker is represented by a 1-layer network.

4. Experiments

In this section, we first describe the experimental settings. We then provide quantitative results that compare our method with state-of-the-art methods in supervised video summarization task. We provide ablation studies with hyperparameters tuning, and demonstrate the relevance of local and personalized summarization. Lastly, we evaluate our results qualitatively with visualization and a user study.

4.1. Datasets Preparation

TVSum [49] dataset is a collection of 50 YouTube videos grouped into 10 categories. Each video is split into a set of 2 second-long shots. 20 users are asked to rate how important each shot is, compared to other shots from the same video in order to build 20 reference summaries. The GT summary for each video is defined as the mean of the corresponding 20 reference summaries.

SumMe [14] dataset is constituted of 25 videos containing a variety of events. For each video, 15 to 18 reference interval-based keyshot summaries were associated. These summaries are converted to frame-level reference summaries by marking the frames contained in the keyshots with score 1 and frames not contained in the keyshots by score 0. Then, the GT summary associated with each video is defined as the mean of 15 to 18 reference summaries.

Methods	TVSum	SumMe	FineGym
Human baseline	0.1755 ± 0.0227	0.1796 ± 0.0107	-
VASNet [11]	0.1690 ± 0.0189	0.0224 ± 0.0289	0.3739 ± 0.0295
dppLSTM [67]	0.0298 ± 0.0284	-0.0256 ± 0.0214	-0.0267 ± 0.0075
DR-DSN$_{60}$ [73]	0.0169 ± 0.0508	**0.0433 ± 0.0386**	0.1457 ± 0.1108
DR-DSN$_{2000}$ [73]	0.1516 ± 0.0373	-0.0159 ± 0.0305	NaN
CSNet+GL+RPE [23]	0.0700 ± 0.0000	-	-
SumGraph [39]	0.0940 ± 0.0000	-	-
Standard ranker	**0.1758 ± 0.0243**	0.0108 ± 0.0407	0.3792 ± 0.0335
Multi-ranker$_8$	0.1750 ± 0.0296	-0.0097 ± 0.0405	-
Multi-ranker$_4$	0.1736 ± 0.0266	-0.0006 ± 0.0454	**0.3928 ± 0.0291**
Multi-ranker$_2$	0.1630 ± 0.0209	0.0172 ± 0.0198	-

Table 1. The mean and standard deviation of Kendall's τ coefficient [24] per each method and dataset. Multi-ranker$_P$ denotes the trained model with P preferences $\mathcal{P} = \{1 \cdots P\}$ and DR-DSN$_{ep}$ denotes the trained model for ep epochs. The best performing model is highlighted and the symbol '-' means that the results are not available. We note that FineGym has only 4 fixed preferences and 1 reference summary.

Unlike TVSum and SumMe that are video summarization datasets, FineGym [45] is a fine-grained action recognition dataset that provides action level temporal annotations for 156 YouTube gymnasium videos. Since the videos are of long duration, we only used 50 sampled videos for experiments purpose and listed their ID in the Supplemental Material. In this case, we do not have reference summaries instead, we define one reference summary and the GT summary for each video by marking the frames contained in the action keyshots with score 1 and frames not contained in the keyshots by score 0.

4.2. Evaluation Metric

The common evaluation metric in the state-of-the-art video summarization methods is F1 score calculated between the predicted and reference summaries where the summarization pipeline consists of importance score estimation, video segmentation and keyshots selection. In recent work, [36] showed that randomly generated summaries achieve similar or better results than the state-of-the-art methods which imply that the importance score estimation part has no major influence on the measure score. Instead, [36] proposed alternative evaluation metrics that compare the importance scores ordering of the reference and predicted summary. These metrics are rank correlation coefficients, precisely Kendall's τ [24] and Spearman's ρ [74] coefficients. In this work, we focus on using Kendall's τ rank correlation coefficient to evaluate our method and compare it with state-of-the-art methods, while we report the results using Spearman's ρ in the Supplemental Material.

4.3. Implementation Details

In FineGym dataset, we only define 4 categorical preferences using action labels (*Vaulting*, *Floor Exercise*, *Balance Beam*, *Uneven Bars*). Segments without annotations are assigned to an additional *background* preference, which is not used in the experiments due to its application irrelevance. In TVSum and SumMe datasets, there is no segment labelling to define as preferences. Instead, we opt for training k-means models on 5000 randomly sampled segment features with $2, 4, 8$ clusters respectively, where the number of clusters represents the number of preferences and each video segment is labelled with the predicted preference.

We generated segment features s_i using 3D ResNet [16] with ResNet-50 [17] backbone pretrained on Kinetics dataset [7]. The features are of 2048 dimensions extracted after the flattening of the pooling of the last conv layer. We note that each feature represents a segment s_i of 16 frames such that its preference p_i is defined as mentioned in the previous paragraph and its corresponding importance score l_i is the mean of the 16 frames GT importance scores resulting in segment-level GT importance scores.

We modelled the ranker R and each sub-ranker $\{R_i\}$ using one fully-connected layer (FC) and trained the Standard ranker and Multi-ranker methods using Adam optimizer [28]. We set the learning rate to 0.0002 in all experiments and the hyperparameter λ, mini-batch size B, number of pairwise comparisons N and training epochs are specified in each experiment according to the Ablation Study 4.5.2.

4.4. Experimental Protocol

Our model is trained using segment-level GT summaries, while the testing of our model and the baseline methods is performed using the frame-level reference summaries. In case a model is trained on segment-level features, the predicted frame importance score is equal to the predicted importance score of the segment containing that frame. Given a video's reference summaries and predicted summary, the resulting video correlation coefficient is the mean of the correlation coefficients between the predicted summary and each reference summary. Additionally, the human baseline correlation coefficient of a video is defined using leave-one-out approach [36], which is the mean of coefficients between each possible pair of reference summaries. We note

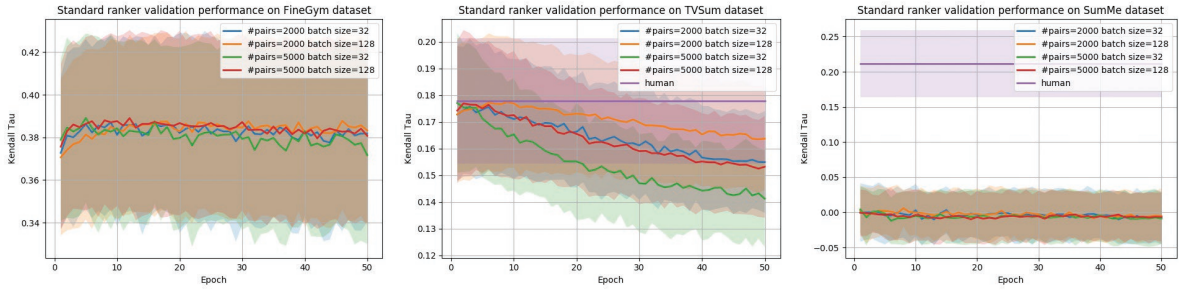

Figure 3. The mean Kendall's τ coefficient per each setting and dataset (solid line) surrounded by a shaded area of range $[-std, std]$, with std the Kendall's τ coefficient standard deviation and $\#pairs$ the number of pairwise comparisons.

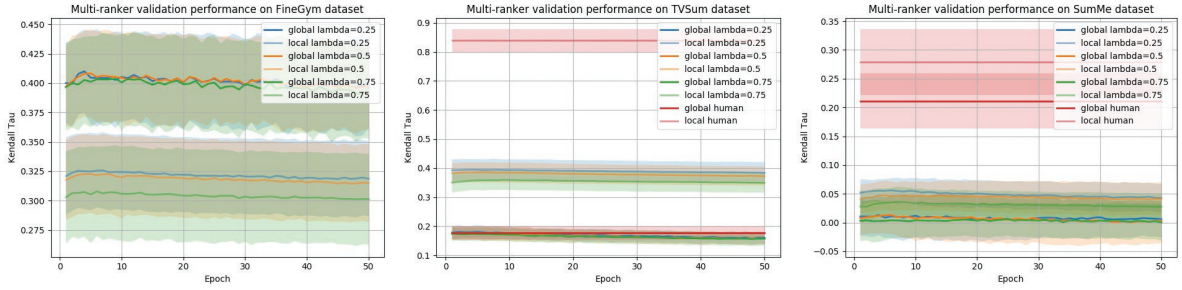

Figure 4. The mean Kendall's τ coefficient per each setting and dataset (solid line) surrounded by a shaded area of range $[-std, std]$, with std the Kendall's τ coefficient standard deviation and *global*, *local* denoting global and local summarization coefficients.

that FineGym dataset has one reference summary and thus the human baseline is not defined. Similarly to [11] canonical setting, we generate 5 random non-test/test splits for each individual dataset. We set 80% of all videos in non-test set and 20% in test set and define the correlation coefficient of a set of videos as the mean of correlation coefficients of each video.

4.5. Quantitative Results

4.5.1 Comparison with State-of-the-art Methods

Baselines: We train dppLSTM [67], VASNet [11], DR-DSN [73], Standard ranker, Multi-ranker on TVSum and SumMe using the feature embeddings described in [11]. For FineGym, we train these models using our feature processing described in Subsection 4.3. In the case of Multi-ranker $\{R_i\}$ and Standard ranker R, we set $N = 2000, B = 128, \lambda = 0.5$ and train them for 1 epoch according to Ablation Study 4.5.2 findings, while the human baseline is defined as in Subsection 4.4. Additionally, we report CSNet+GL+RPE [23] and SumGraph [39] original results on TVSum since no implementation is publicly available. We note that unless mentioned otherwise, all these models are trained and tested on the same sets using their default hyperparameters.

Following the Experimental Protocol 4.4, we compare our Multi-ranker with these baselines on the global summarization task using the test set of each split in each benchmark. We report in Table 1, the mean and standard deviation of Kendall's τ coefficients on the test sets.

On TVSum dataset, we notice that Standard ranker achieves comparable results to the human baseline with Multi-ranker and VASNet slightly below. dppLSTM and DR-DSN struggle to generalize to test sets, however when we trained DR-DSN for more epochs, it performs well comparable to other methods. Similarly to TVSum, on FineGym dataset, Standard ranker and VASNet perform well with Multi-ranker slightly above. dppLSTM struggles to generalize to test sets, while DR-DSN showed unstable predictions and eventually with more training epochs, it diverges by predicting the same score for all segments.

Concerning SumMe dataset, we could not neglect the fact that all experimented methods have failed to generalize on the test set. So far, the previous works that relied on the rank correlation coefficient evaluation [36, 23, 39] have only shown results on TVSum. [23] mentioned that the binary importance scores in SumMe reference summaries are not a proper form for the evaluation metrics. We disagree with this proposition while providing the performance on FineGym dataset with the binary importance scores in its reference summaries as an evidence. Evaluating with this reference summary form will induce many pairwise ties that need to be accounted for by the rank correlation measure. Fortunately, Kendall's Tau-b and Spearman's Rho statistics have adjustments for ties [2, 74] and can be safely used. On the other hand, we make the observations that SumMe videos are widely context independent while FineGym videos have the same context, and also that our method and the baselines have failed to generalize

using two different video feature extractors (more details in Supplemental Material). These observations lead us to question the generalization capability of the training segment features of SumMe videos.

4.5.2 Ablation Study

We conduct ablation studies to tune the hyperparameters of our model and investigate their impact on performance. Since an exhaustive grid search is time consuming, we use the following two-step process: we first tune the mini-batch size B and number of pairwise comparisons N by training a Standard ranker R, then we tune the hyperparameter λ by training a Multi-ranker $\{R_i\}$.

We set $B \in \{32, 128\}$, $N \in \{2000, 5000\}$ and follow the Experimental Protocol 4.4 in training Standard ranker R using 4-fold cross-validation on the non-test set for each split. As a result, we train 20 models for 50 epochs and report in Figure 3 the mean and standard deviation of Kendall's τ coefficients on the validation sets along with the corresponding human baseline. According to the plots in Figure 3, we made the following conclusions: the early epochs are enough to obtain an optimal ranker and further training leads to overfitting on training set. The model is not sensitive to the hyperparameters B and N at optimal epochs since the coefficient differences are not significant.

The aim for hyperparameter λ is to balance the local and global summarization of Multi-ranker. In order to quantify this trade-off, reporting correlation coefficients related to predicted global summaries is not enough. We define a local reference summary for a video with respect to preference j such that a frame importance score is set to 0 if the associated segment s_i satisfies $p_i \neq j$. Thus, the resulting video local correlation coefficient is the mean of the correlation coefficients between the predicted and reference local summary with respect to each local reference summary and each preference. Also, the local human baseline correlation coefficient is the mean of coefficients between each possible pair of local reference summaries for each preference.

We set $\mathcal{P} = \{1 \dots 4\}$, $B = 128$, $N = 2000$, $\lambda \in \{0.25, 0.5, 0.75\}$ and follow the Experimental Protocol 4.4 in training Multi-ranker $\{R_i\}$ using 4-fold cross validation on the non-test set for each split. We report in Figure 4 the mean and standard deviation of local and global Kendall's τ coefficients of the validation sets along with the corresponding human baselines. We notice that the λ variation does not have an impact on the global summarization performance, while the mean local correlation coefficient tend to decrease when λ puts more emphasis on global summarization. In addition, the influence of the number of preferences on the global summarization is also investigated and shown in the Supplemental Material.

Pref. set	Multi-ranker	Standard ranker
{1}	0.1086 ± 0.0164	0.0254 ± 0.0122
{2}	0.3568 ± 0.0376	0.2727 ± 0.0241
{3}	0.3985 ± 0.0097	0.2978 ± 0.0133
{4}	0.3007 ± 0.0283	0.1504 ± 0.0840
{1,2}	0.3928 ± 0.0291	0.3792 ± 0.0335
{1,3}	0.3747 ± 0.0245	0.2829 ± 0.0325
{1,4}	0.2359 ± 0.0286	0.1200 ± 0.0582
{2,3}	0.4093 ± 0.0135	0.3925 ± 0.0183
{2,4}	0.3707 ± 0.0218	0.2781 ± 0.0387
{3,4}	0.3966 ± 0.0117	0.2996 ± 0.0201
{1,2,3}	0.3928 ± 0.0291	0.3792 ± 0.0335
{1,2,4}	0.3928 ± 0.0291	0.3792 ± 0.0335
{1,3,4}	0.3747 ± 0.0245	0.2829 ± 0.0325
{2,3,4}	0.4093 ± 0.0135	0.3925 ± 0.0183
{1,2,3,4}	0.3928 ± 0.0291	0.3792 ± 0.0335

Table 2. The mean and standard deviation Kendall's τ coefficient of Multi-ranker and Standard ranker for each possible preference set $\mathcal{P}_s$ (Pref. set).

4.5.3 Relevance of Personalized Summarizations

This experiment's aim is to demonstrate that Multi-ranker provides more preference specific summaries than the Standard ranker. For this purpose, we define a personalized reference summary for a video with respect to preference set $\mathcal{P}_s$ such that a frame importance score is set to 0 if the associated segment s_i satisfies $p_i \notin \mathcal{P}_s$. Thus, the resulting video personalized correlation coefficient is the mean of the correlation coefficient between the predicted personalized summary and each personalized reference summary. Although the Standard ranker is trained on GT summaries to generate a global summary, testing it using personalized reference summaries sets a lower bound baseline for Multi-ranker.

We set $N = 2000, B = 128, \lambda = 0.5, \mathcal{P} = \{1 \dots 4\}$ and train Multi-ranker and Standard ranker for 1 epoch. Following the Experimental Protocol 4.4, we test these models on the personalized summarization task using the test set of each split in FineGym dataset. We report in Table 2, the mean and standard deviation of personalized Kendall's τ coefficients on the test sets. As expected from Multi-ranker model, we notice that the more general the generated summary is, the more the Multi-ranker correlation coefficient is similar to the Standard ranker. Also, the more local the generated summary is, the wider the disparity between Standard ranker and Multi-ranker correlation coefficients.

4.6. Qualitative Results

4.6.1 Visualization

In this subsection, we present an example of global and local video summaries in FineGym dataset, with more examples shown in the Supplemental Material. In Figure 5, we illustrate the segment-level GT importance scores of a video in FineGym and highlight the top-ranked 15% global and local predicted segments with respect to *Floor Exercise*

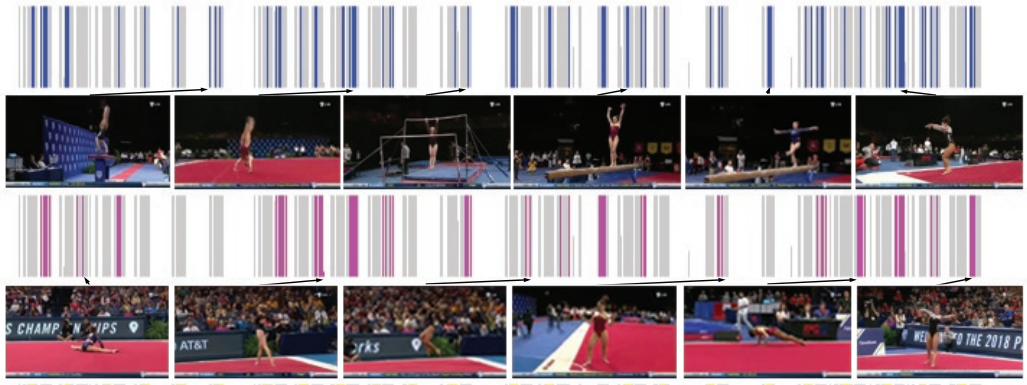

Figure 5. Segment-level GT importance scores (gray), Multi-ranker global summary (blue), Multi-ranker local summary for *Floor Exercise* preference (magenta) for the test video with ID '0LtLS9wROrk' from FineGym dataset.

preference using Multi-ranker in local and global summarization tasks. In each illustrated summary, we visualize 6 sampled frames from the highlighted predicted frames.

4.6.2 User Study

To quantify the perceived quality of our Multi-ranker method and the impact from the user perspective of each Multi-ranker task, *i.e.* local, personalized and global summarizations, we performed a user study based on 40 subjects that are asked to provide their opinions about 4 main comparison scenarios. We focused on FineGym videos with its predefined 4 preferences and asked the subject to watch for each scenario run the original video and two associated summaries while selecting a preference for local summary or a set of preferences for personalized summary and then submit his answer to the scenario question. The first scenario is a subjective comparison between Multi-ranker and VASNet [11] summaries, while the remaining scenarios are comparisons between local, personalized and global summaries in term of usability and satisfaction from the user perspective. The corresponding question for each scenario is defined as follows: (1) Is the quality of Multi-ranker summary better, equal or worse than VASNet [11] summary? (2) Is local summary more content specific than global summary or not? (3) Does personalized summary provide better user control to achieve satisfactory result than global summary or not? (4) Does personalized summary provide better user control to achieve satisfactory result than local summary or not?

Table 3 shows the user study results. In short, nearly half of the participants found that Multi-ranker and VASNet [11] summaries have similar quality, which is compatible with the quantitative comparison. The majority of participants found that local summary is more content specific than global summary and that personalized summary has better user control and satisfaction than local and global summaries. These results are a promising indication that user interactive summarization is more appealing and satisfying than unique global summarization.

	SD	MD	Similar	MA	SA
Scen. 1	7.14%	14.29%	**46.43%**	21.43%	10.71%
Scen. 2	0.00%	5.49%	5.49%	24.18%	**64.84%**
Scen. 3	2.56%	6.41%	10.26%	20.51%	**60.26%**
Scen. 4	0.00%	3.85%	2.56%	**85.90%**	7.69%

Table 3. User study results with respect to each scenario (Scen.) with (SD,MD,MA,SA) stands for (Strongly Disagree, Mildly Disagree, Mildly Agree, Strongly Agree) respectively. The numbers indicate the percentage of responses for each scenario question.

5. Discussion and Conclusion

We introduce Multi-ranker, a multiple pairwise ranking model for personalized video summarization using predefined preferences. We proposed a training scheme allowing the model to accomplish local, personalized, or global summarization tasks. Our experiments show that the proposed method can generate not only high-quality global summaries that are comparable to the state-of-the-art, but also personalized summaries that conform with a set of preferences. We believe that a proper benchmark such as a sport summarization dataset is needed to explore the range of possible applications in the video summarization tasks. In this work, we only focused on the adaptability to the user preference summarization criteria and further criteria such as the diversity and coherence can be explored as future works.

6. Acknowledgements

This work is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 665992, CDE - the UK's EPSRC Centre for Doctoral Training in Digital Entertainment (EP/L016540/1), RCUK grant CAMERA (EP/M023281/1, EP/T022523/1), and a gift from Adobe. We thank the Alibaba Group for supporting the User Study.

References

[1] Lalitha Agnihotri, John R. Kender, Nevenka Dimitrova, and John Zimmerman. Framework for personalized multimedia summarization. In *MIR*, 2005. 2

[2] Alan Agresti. *Analysis of ordinal categorical data*. John Wiley & Sons, 2010. 6

[3] Evlampios E. Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks: A survey. *CoRR*, 2021. 2

[4] Noboru Babaguchi, Kouzou Ohara, and Takehiro Ogura. Learning personal preference from viewer's operations for browsing and its application to baseball video retrieval and summarization. *IEEE Trans. Multim.*, 2007. 2

[5] Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *ICML*, 2005. 3

[6] Sijia Cai, Wangmeng Zuo, Larry S. Davis, and Lei Zhang. Weakly-supervised video summarization using variational encoder-decoder and web prior. In *ECCV*, 2018. 2

[7] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 5

[8] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015. 2

[9] Ana Garcia del Molino, Xavier Boix, Joo-Hwee Lim, and Ah-Hwee Tan. Active video summarization: Customized summaries via on-line interaction with the user. In *AAAI*, 2017. 3

[10] Ana Garcia del Molino and Michael Gygli. PHD-GIFs: Personalized highlight detection for automatic GIF creation. In *ACM Multimedia*, 2018. 1, 3, 4

[11] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *ACCV*, 2018. 1, 2, 5, 6, 8

[12] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *NeurIPS*, 2014. 1, 2

[13] Michael Gygli, Helmut Grabner, and Luc Van Gool. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 2015. 1, 2

[14] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*, 2014. 2, 4

[15] Michael Gygli, Yale Song, and Liangliang Cao. Video2GIF: Automatic generation of animated GIFs from video. In *CVPR*, 2016. 1, 2, 3

[16] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? In *CVPR*, 2018. 5

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[18] Xufeng He, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Martin Robertson, and Haibing Guan. Unsupervised video summarization with attentive conditional generative adversarial networks. In *ACM Multimedia*, 2019. 2

[19] Fa-Ting Hong, Xuanteng Huang, Weihong Li, and Wei-Shi Zheng. MINI-Net: Multiple instance ranking network for video highlight detection. In *ECCV*, 2020. 1, 2

[20] Alejandro Jaimes, Tomio Echigo, Masayoshi Teraguchi, and Fumiko Satoh. Learning personalized video highlights from detailed MPEG-7 metadata. In *ICIP*, 2002. 2

[21] Yifan Jiao, Xiaoshan Yang, Tianzhu Zhang, Shucheng Huang, and Changsheng Xu. Video highlight detection via deep ranking modeling. In *PSIVT*, 2017. 2

[22] Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. Discriminative feature learning for unsupervised video summarization. In *AAAI*, 2019. 2

[23] Yunjae Jung, Donghyeon Cho, Sanghyun Woo, and In So Kweon. Global-and-local relative position embedding for unsupervised video summarization. In *ECCV*, 2020. 1, 2, 5, 6

[24] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. 5

[25] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013. 1, 2

[26] Gunhee Kim, Leonid Sigal, and Eric P. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*, 2014. 2

[27] Gunhee Kim and Eric P. Xing. Reconstructing storyline graphs for image recommendation from web community photos. In *CVPR*, 2014. 2

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[29] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2

[30] Zutong Li and Lei Yang. Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward. In *WACV*, 2021. 2

[31] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*, 2015. 2

[32] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 2

[33] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial LSTM networks. In *CVPR*, 2017. 1, 2

[34] Engin Mendi, Hélio B. Clemente, and Coskun Bayrak. Sports video summarization based on motion analysis. *Comput. Electr. Eng.*, 2013. 2

[35] Chong-Wah Ngo, Yu-Fei Ma, and HongJiang Zhang. Automatic video summarization by graph modeling. In *ICCV*, 2003. 2

[36] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Rethinking the evaluation of video summaries. In *CVPR*, 2019. 5, 6

[37] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K. Roy-Chowdhury. Weakly supervised summarization of web videos. In *ICCV*, 2017. 1, 2

[38] Rameswar Panda and Amit K. Roy-Chowdhury. Collaborative summarization of topic-related videos. In *CVPR*, 2017. 1, 2

[39] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. SumGraph: Video summarization via recursive graph modeling. In *ECCV*, 2020. 1, 2, 5, 6

[40] Danila Potapov, Matthijs Douze, Zaïd Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *ECCV*, 2014. 2

[41] Mrigank Rochan, Mahesh Kumar Krishna Reddy, Linwei Ye, and Yang Wang. Adaptive video highlight detection by learning from user history. In *ECCV*, 2020. 1, 3, 4

[42] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *CVPR*, 2019. 1, 2

[43] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *ECCV*, 2018. 2

[44] Yong Rui, Anoop Gupta, and Alex Acero. Automatically extracting highlights for TV baseball programs. In *ACM Multimedia*, 2000. 2

[45] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. FineGym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, 2020. 5

[46] Aidean Sharghi, Boqing Gong, and Mubarak Shah. Query-focused extractive video summarization. In *ECCV*, 2016. 1, 3

[47] Aidean Sharghi, Jacob S. Laurel, and Boqing Gong. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *CVPR*, 2017. 3

[48] Mohammad Soleymani. The quest for visual interest. In *ACM Multimedia*, 2015. 1

[49] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. TVSum: Summarizing web videos using titles. In *CVPR*, 2015. 2, 4

[50] Yaser Souri, Erfan Noury, and Ehsan Adeli. Deep relative attributes. In *ACCV*, 2016. 3

[51] Min Sun, Ali Farhadi, and Steven M. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *ECCV*, 2014. 1, 2

[52] Hao Tang, Vivek Kwatra, Mehmet Emre Sargin, and Ullas Gargi. Detecting highlights in sports videos: Cricket as a test case. In *ICME*, 2011. 2

[53] Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multim. Comput. Commun. Appl.*, 2007. 1

[54] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Query-adaptive video summarization via quality-aware relevance estimation. In *ACM Multimedia*, 2017. 1, 3

[55] Jinjun Wang, Changsheng Xu, Chng Eng Siong, and Qi Tian. Sports highlight detection from keyword sequences using HMM. In *ICME*, 2004. 2

[56] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. Convolutional hierarchical attention network for query-focused video summarization. In *AAAI*, 2020. 3

[57] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *CVPR*, 2019. 1, 2, 3

[58] Bo Xiong, Gunhee Kim, and Leonid Sigal. Storyline representation of egocentric videos with an applications to story-based search. In *ICCV*, 2015. 2

[59] Ziyou Xiong, Regunathan Radhakrishnan, Ajay Divakaran, and Thomas S. Huang. Highlights extraction from sports video based on an audio-visual marker detection framework. In *ICME*, 2005. 2

[60] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M. Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *CVPR*, 2015. 2

[61] Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. VideoQA: question answering on news video. In *ACM Multimedia*, 2003. 2

[62] Huan Yang, Baoyuan Wang, Stephen Lin, David P. Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *ICCV*, 2015. 1, 2

[63] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*, 2016. 2, 3

[64] Youngjae Yu, Sangho Lee, Joonil Na, Jaeyun Kang, and Gunhee Kim. A deep ranking model for spatio-temporal highlight detection from a 360° video. In *AAAI*, 2018. 2

[65] Li Yuan, Francis E. H. Tay, Ping Li, Li Zhou, and Jiashi Feng. Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization. In *AAAI*, 2019. 2

[66] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *CVPR*, 2016. 2

[67] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, 2016. 1, 2, 5, 6

[68] Ke Zhang, Kristen Grauman, and Fei Sha. Retrospective encoders for video summarization. In *ECCV*, 2018. 2

[69] Yujia Zhang, Michael C. Kampffmeyer, Xiaodan Liang, Min Tan, and Eric P. Xing. Query-conditioned three-player adversarial network for video summarization. In *BMVC*, 2018. 3

[70] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hierarchical recurrent neural network for video summarization. In *ACM Multimedia*, 2017. 2

[71] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. HSA-RNN: hierarchical structure-adaptive RNN for video summarization. In *CVPR*, 2018. 2

[72] Bin Zhao and Eric P. Xing. Quasi real-time summarization for consumer videos. In *CVPR*, 2014. 2

[73] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *AAAI*, 2018. 1, 2, 5, 6

[74] Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. CRC Press, 1999. 5, 6