Multimodal Frame-Scoring Transformer for Video Summarization

Jeiyoon Park[1,3], **Kiho Kwoun**[2], **Chanhee Lee**[4], **Heuiseok Lim**[1]
[1]Department of Computer Science and Engineering, Korea University
[2]School of Electrical and Computer Engineering, University of Seoul
[3]LLSOLLU
[4]Naver Corporation

Abstract

As the number of video content has mushroomed in recent years, automatic video summarization has come useful when we want to just peek at the content of the video. However, there are two underlying limitations in generic video summarization task. First, most previous approaches read in just visual features as input, leaving other modality features behind. Second, existing datasets for generic video summarization are relatively insufficient to train a caption generator and multimodal feature extractors. To address these two problems, this paper proposes the Multimodal Frame-Scoring Transformer (MFST) framework exploiting visual, text, and audio features and scoring a video with respect to frames. Our MFST framework first extracts each modality features (visual-text-audio) using pretrained encoders. Then, MFST trains the multimodal frame-scoring transformer that uses video-text-audio representations as inputs and predicts frame-level scores. Our extensive experiments with previous models and ablation studies on TVSum and SumMe datasets demonstrate the effectiveness and superiority of our proposed method.

1 Introduction

When humans watch a video on YouTube or Netflix, they perceive visual, linguistic, and audio information through various sense organs and know which parts of the video are interesting. To consider whether a scene in a movie is absorbing, for example, we observe characters' facial expressions and actions, recognize background and situation with language, and listen to the characters' utterances and sound effects. Intuitively, humans have access to well-defined scoring function in their mind, using versatile sensory systems, while video summarization models from previous studies did not.

Video summarization aims to capture key frames using predicted frame-wise importance scores, given datasets as shown in Figure 1. However,

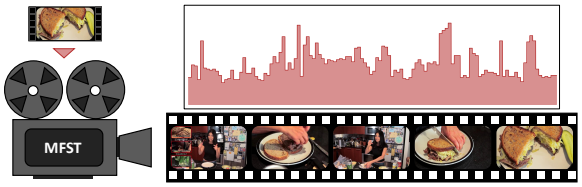

Figure 1: The proposed MFST takes a video as input and predicts importance scores for video summarization. The red areas indicate importance score with respect to frames.

most existing approaches exploit just visual features, leaving other modality features behind (Zhao et al., 2018; Rochan et al., 2018; Zhang et al., 2018; Zhou et al., 2018; Jung et al., 2019; Rochan and Wang, 2019; Park et al., 2020; Jung et al., 2020; Ghauri et al., 2021).

Language-attended methods employ generated video captions (Bor-Chun Chen and Chen, 2017; Narasimhan et al., 2021), or video descriptions (Haopeng et al., 2022) from videos while training. Though language-guided approaches alleviate modality issue somewhat, there are underlying constraints in conveying vivid audio features into a video summarization model (e.g., we know there are limits to expressing a beautiful song just with the text *"beautiful song"*).

This paper proposes Multimodal Frame-Scoring Transfomer (MFST) to handle different modality combinations and frame-level scoring for video summarization. Note that existing datasets for generic video summarization (Gygli et al., 2014; Song et al., 2015) are relatively insufficient to pretrain a dense caption generator and visual, text, and audio feature extractors. We investigate a new multimodal setting where it can mitigate the lack of human-scored videos used for training video summarization model.

To this end, our framework consists of two stages: (i) Extracting each modality features (visual-text-audio) using pretrained encoders. (ii) Multimodal frame-scoring transformer that reads

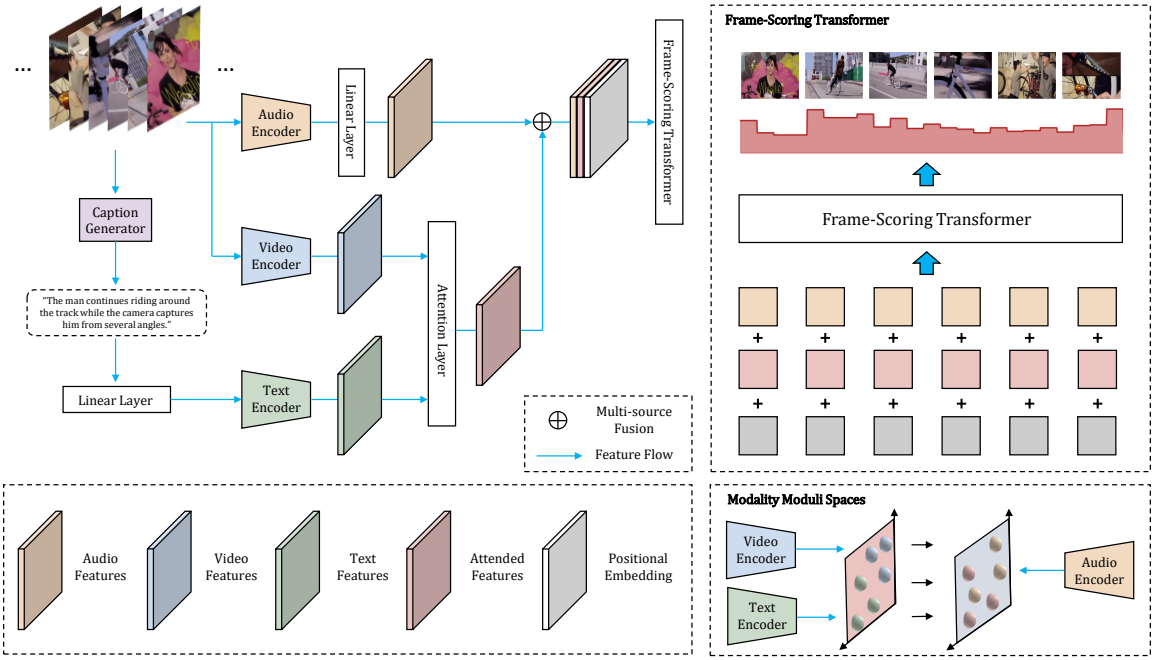

Figure 2: Schematic depiction of the proposed Multimodal Frame-Scoring Transformer (MFST) for video summarization. Given a video, we first generate dense video captions using learned caption generator. To mitigate the lack of human-scored videos, we extract each modality features (visual-text-audio) from the video and generated captions, exploiting learned VideoCLIP and Wav2Vec2. Then, we calculate text-attended visual representation using attention layer and compute Multi-source fusion across the text-attended features and audio features. Finally we feed the fused representation to frame-scoring transformer with positional encodings at the bottom of transformer encoder and decoder stacks. We modify the transformer so that it can use modality-fused representations as input and predict frame-level scores.

in video-text-audio representations as inputs and predicts frame-level scores.

Our extensive experiments with previous models and ablation studies on TVSum and SumMe datasets demonstrate the effectiveness and superiority of our proposed method.

Contributions.

1. To the best of our knowledge, Our MFST is the first to introduce frame-scoring transformer exploiting multimodal features (visual-text-audio) for generic video summarization task.
2. We investigate a new multimodal setting where it can mitigate the lack of human-scored videos used for training generic video summarization model using pretrained modules.
3. Our empirical study on generic video summarization datasets (TVSum and SumMe) demonstrates the effectiveness and superiority of our approach.

Related Work. In video summarization task, there are two broad categories of methods: (i) query-guided video summarization (Narasimhan et al., 2021; Wu et al., 2022; Liu et al., 2022; Jiang and Mu, 2022) and (ii) generic video summarization (Park et al., 2020; Jung et al., 2020; Ghauri et al., 2021; Narasimhan et al., 2021; Haopeng et al., 2022). The first category of methods find relevant moments according to user-defined query. Note that though query-driven approach is necessary because defining salient scenes is often subjective, it is difficult to apply if we do not know the contents of the video or if we do not need a subjective summary (e.g., YouTube video previews).

In this paper, we aim to tackle the second category based on a novel multimodal frame-scoring framework.

2 Preliminaries

We consider the standard generic video summarization setting, where given a set of videos V and ground truth frame scores S_{gt}, the goal is to minimize the loss $\mathcal{L}_\theta$ with respect to predicted frame scores S:

$$\hat{\theta} = \arg\min_{\theta}(\mathcal{L}_\theta(S_{gt}, S)) \quad (1)$$

Note that existing datasets for generic video summarization (Gygli et al., 2014; Song et al., 2015)

are relatively insufficient to train a caption generator and visual, text, and audio feature extractors. Considering that large-scaled videos annotated by humans are not available, we exploit the pretrained caption generator and feature extractor for each modality: $C = g_c(V), \mathcal{T} = f_t(C), \mathcal{V} = f_v(V), \mathcal{A} = f_a(V)$, where C denotes dense video captions, g_c is a dense video caption generator (Iashin and Rahtu, 2020), f_t is a text feature extractor (Xu et al., 2021), f_v is a visual feature extractor (Xu et al., 2021), and f_a is a audio feature extractor (Baevski et al., 2020).

3 Approach

In this section, we present the proposed MFST framework (as shown in Figure 2) with extracted modality features described in the previous section.

3.1 Fine and Coarse Moduli Spaces

To represent text-attended visual representation, we consider modality moduli spaces. Inspired by (Alayrac et al., 2020), but unlike this study, we first put a fine moduli space where visual features and text features lie:

$$\Psi_M : \mathcal{V} \rightarrow h_{\mathcal{V}} \tag{2}$$

$$\Psi_M : \mathcal{T} \rightarrow h_{\mathcal{T}} \tag{3}$$

where, Ψ_M is a contravariant functor, $\mathcal{V}$ is visual features and $\mathcal{T}$ is text features. Note that (Liang et al., 2022) demonstrates "multimodal video-text pretraining" paradigm can not solve the *modality gap* phenomenon completely which causes performance degradation. In this work, we compute text-attended visual representation using attention layer:

$$h_{\mathcal{VT}} = \text{Attention}(h_{\mathcal{V}}, h_{\mathcal{T}}) \tag{4}$$

where, $h_{\mathcal{VT}}$ denotes fine moduli space. Then, we project the fine moduli space into the coarse moduli space by modality fusion:

$$h_{\mathcal{VTA}} = \mathcal{F}_{M'}(h_{\mathcal{VT}}, h_{\mathcal{A}}) \tag{5}$$

where, $\mathcal{F}_{M'}$ denotes function of fusion and $h_{\mathcal{VTA}}$ represents coarse moduli space.

3.2 Multimodal Frame-Scoring Transformer

In this paper, we introduce frame-scoring transformer to video summarization, which is modified to predict frame-level importance scores S. We feed $h_{\mathcal{VTA}}$ to frame-scoring transformer with positional encodings at the bottom of transformer encoder and decoder stacks. Given ground truth frame scores S_{gt} of N frames from a video, we train MFST using the mean square error:

$$\mathcal{L}_\theta(S_{gt}, S) = \frac{1}{N} \left\| S_{gt} - S \right\|_2^2 . \tag{6}$$

Table 1: Experimental results on SumMe and TVSum under the Canonical, Augment, and Transfer settings (F-score).

Methods	SumMe			TVSum		
	Can	*Aug*	*Tran*	*Can*	*Aug*	*Tran*
vsLSTM (Zhang et al., 2016)	0.376	0.416	0.407	0.542	0.579	0.569
SGAN (Mahasseni et al., 2017)	0.387	0.417	—	0.508	0.589	—
SGAN$_s$ (Mahasseni et al., 2017)	0.417	0.436	—	0.563	0.612	—
H-RNN (Zhao et al., 2017)	0.421	0.438	—	0.579	0.619	—
DRDSN (Zhou et al., 2018)	0.421	0.439	0.426	0.581	0.598	0.589
HSA-RNN (Zhao et al., 2018)	0.423	0.421	—	0.587	0.598	—
ACGAN (He et al., 2019)	0.460	0.470	0.445	0.585	0.589	0.578
WS-HRL (Chen et al., 2019)	0.436	0.445	—	0.584	0.585	—
re-S2S (Zhang et al., 2018)	0.425	0.449	—	0.603	0.639	—
S-FCN (Rochan et al., 2018)	0.475	0.511	0.441	0.568	0.592	0.582
VASNet (Fajtl et al., 2018)	0.497	0.510	—	0.614	0.623	—
CSNet$_s$ (Jung et al., 2019)	0.513	0.521	0.451	0.588	0.590	0.592
GLRPE (Jung et al., 2020)	0.502	—	—	0.591	—	—
SumGraph (Park et al., 2020)	0.514	0.529	0.487	0.639	0.658	0.605
RSGN (Zhao et al., 2021)	0.450	0.457	0.440	0.601	0.611	0.600
RSGN$_{uns}$ (Zhao et al., 2021)	0.423	0.436	0.412	0.580	0.591	0.597
MSVA (Ghauri et al., 2021)	0.545	—	—	0.628	—	—
CLIP-It (Narasimhan et al., 2021)	0.542	0.564	0.519	0.663	0.690	0.655
SSPVS (Haopeng et al., 2022)	0.501	—	—	0.607	—	—
iPTNet (Jiang and Mu, 2022)	0.545	0.569	0.492	0.634	0.642	0.598
MFST (Ours)	**0.595**	**0.655**	**0.576**	**0.737**	**0.779**	**0.691**

Table 2: Experimental results on SumMe and TVSum (Kendall's τ and Spearman's ρ).

Methods	SumMe		TVSum	
	τ	ρ	τ	ρ
Random	0.000	0.000	0.000	0.000
Human	0.205	0.213	0.177	0.204
Ground Truth	1.000	1.000	0.364	0.456
SGAN (Mahasseni et al., 2017)	—	—	0.024	0.032
WS-HRL (Chen et al., 2019)	—	—	0.078	0.116
DRDSN (Zhou et al., 2018)	0.047	0.048	0.020	0.026
dppLSTM (Zhang et al., 2016)	—	—	0.042	0.055
CSNet$_s$ (Jung et al., 2019)	—	—	0.025	0.034
GLRPE (Jung et al., 2020)	—	—	0.070	0.091
HSA-RNN (Zhao et al., 2018)	0.064	0.066	0.082	0.088
RSGN (Zhao et al., 2021)	0.083	0.085	0.083	0.090
RSGN$_u$ (Zhao et al., 2021)	0.071	0.073	0.048	0.052
SumGraph (Park et al., 2020)	—	—	0.094	0.138
SSPVS (Haopeng et al., 2022)	0.123	0.170	0.169	**0.231**
MSVA (Ghauri et al., 2021)	0.200	**0.230**	0.190	0.210
MFST (Ours)	**0.229**	0.229	**0.222**	0.224

4 Experiments

4.1 Dataset Description

We conduct our video summarization experiments on two benchmarks: TVSum (Song et al., 2015) and SumMe (Gygli et al., 2014). TVSum contains 50 videos, including the topics of news, documentaries. The duration of each video varies from 1 to 10 minutes. 20 annotators provide frame-level importance scores for each video. SumMe consists of 25 user videos, covering various topics (e.g., holidays and sports). Each video ranges from 1

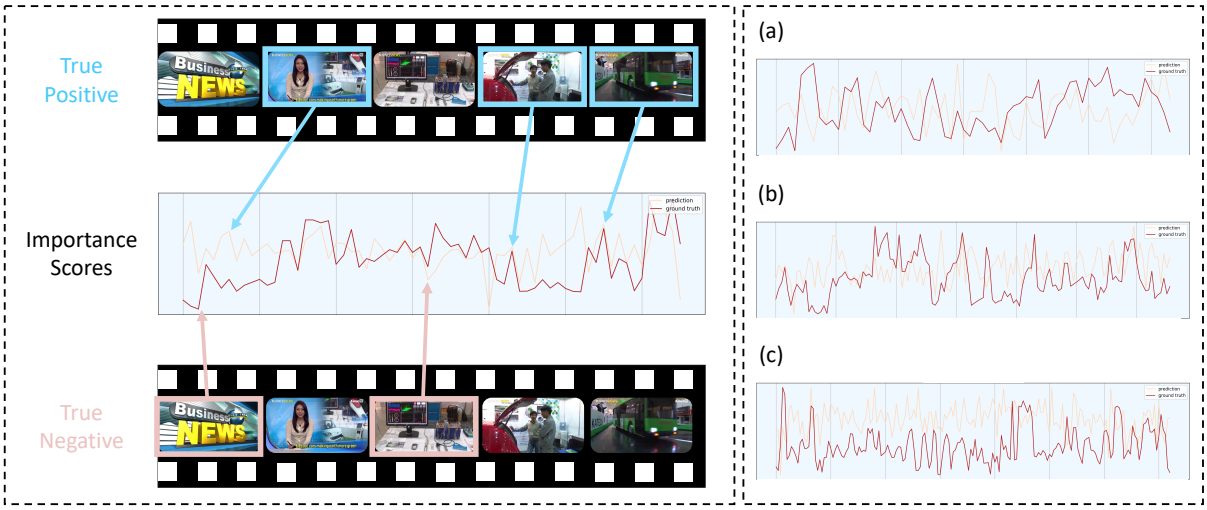

Figure 3: Comparison of predicted score and averaged ground truth scores from TVSum dataset.

Table 3: The results of ablation studies in multimodality.

Methods	SumMe			TVSum		
	Can	*Aug*	*Tran*	*Can*	*Aug*	*Tran*
MFST	0.542	0.629	0.553	0.708	0.753	0.659
MFST + Audio (Ours)	**0.595**	**0.655**	**0.576**	**0.737**	**0.779**	**0.691**

to 6 minutes. 15 to 18 persons annotated multiple ground truth summaries for each video.

4.2 Metric Description

We follow the same experimental metrics used in existing works: F-score and Rank-based evaluation. *True positive* means highlight overlaps between model-generated summary V_m and human-generated summary V_h based on importance scores. The precision and recall are calculated as follows: Precision $= \frac{|V_m \cap V_h|}{|V_m|}$ and Recall $= \frac{|V_m \cap V_h|}{|V_h|}$.

Rank-based evaluations (Otani et al., 2019) compute Kendall's τ and Spearman's ρ which measure non-parametric rank correlations: $\tau = Kendall(S_{gt}, S)$ and $\rho = Spearman(S_{gt}, S)$.

4.3 Performance Comparison

Experimental Settings. We compare MFST with existing models in three different experimental settings: (i) In Canonical setting, we selects the dataset (e.g., TVSum or SumMe) and randomly splits the dataset into training and evaluation. (ii) In Augment setting, we merges the two datasets into one and randomly splits the dataset into training and evaluation. (iii) In Transfer setting, we trains a model using one dataset and evaluates the trained model on the other dataset. Note that in all experimental settings, we conduct experiments over 5 times and average the results. Each experiment randomly selects 20% of the dataset for evaluation.

Results on Video Summarization. Table 1 and Table 2 show our extensive experiments with previous methods on TVSum and SumMe datasets. Under the canonical, augment, and transfer settings, we demonstrate that MFST outperforms existing methods on both benchmarks. It should be noted that the performance of models using two or more modalities is higher than models using unimodality and the results of MFST, which properly exploits the visual, text and audio modalities, demonstrate the effectiveness and superiority of our approach.

Qualitative Results. Figure 3 shows qualitative results on generated summary from MFST with human-generated summary. Note that model-generated score graph and ground truth graph are very similar, which means our model predicts parts that humans find interesting or not.

4.4 Ablation Studies

We further conduct ablation studies as shown in Table 3. To validate the effectiveness of our audio-attended representation, we compare two results, the one without audio features and the other with audio features. Results demonstrate MFST can be advanced when extracted audio features are added.

5 Conclusion

We propose MFST, a novel and effective frame-scoring framework given a video. Unlike existing methods, MFST exploits video-text-audio features using learned feature extractors and frame-scoring multimodal transformer. Our extensive comparisons with previous approaches and ablation studies demonstrate the effectiveness and superiority of our proposed method.

References

Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Yan-Ying Chen Bor-Chun Chen and Francine Chen. 2017. Video to text summary: Joint video summarization and captioning with recurrent neural networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 118.1–118.14. BMVA Press.

Yiyan Chen, Li Tao, Xueting Wang, and Toshihiko Yamasaki. 2019. Weakly supervised video summarization by hierarchical reinforcement learning. In *Proceedings of the ACM Multimedia Asia*, pages 1–6.

Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. Summarizing videos with attention. In *Asian Conference on Computer Vision*, pages 39–54. Springer.

Junaid Ahmed Ghauri, Sherzod Hakimov, and Ralph Ewerth. 2021. Supervised video summarization via multiple feature sets with parallel attention.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Computer Vision – ECCV 2014*, pages 505–520, Cham. Springer International Publishing.

Li Haopeng, Ke Qiuhong, Gong Mingming, and Zhang Rui. 2022. Video summarization based on video-text modelling.

Xufeng He, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. 2019. Unsupervised video summarization with attentive conditional generative adversarial networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2296–2304.

Vladimir Iashin and Esa Rahtu. 2020. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *British Machine Vision Conference (BMVC)*.

Hao Jiang and Yadong Mu. 2022. Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16388–16398.

Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. 2019. Discriminative feature learning for unsupervised video summarization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8537–8544. AAAI Press.

Yunjae Jung, Donghyeon Cho, Sanghyun Woo, and In So Kweon. 2020. Global-and-local relative position embedding for unsupervised video summarization. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV*, page 167–183, Berlin, Heidelberg. Springer-Verlag.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*.

Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multimodal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211.

Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. 2021. Clip-it! language-guided video summarization. In *Advances in Neural Information Processing Systems*, volume 34, pages 13988–14000. Curran Associates, Inc.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. 2019. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7596–7604.

Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. 2020. Sumgraph: Video summarization via recursive graph modeling. In *ECCV*.

Mrigank Rochan and Yang Wang. 2019. Video summarization by learning from unpaired data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7894–7903.

Mrigank Rochan, Linwei Ye, and Yang Wang. 2018. Video summarization using fully convolutional sequence networks. In *Computer Vision – ECCV 2018*, pages 358–374, Cham. Springer International Publishing.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187.

Guande Wu, Jianzhe Lin, and Claudio T. Silva. 2022. Intentvizor: Towards generic query guided interactive video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10503–10512.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer.

Ke Zhang, Kristen Grauman, and Fei Sha. 2018. Retrospective encoders for video summarization. In *Computer Vision – ECCV 2018*, pages 391–408, Cham. Springer International Publishing.

Bin Zhao, Haopeng Li, Xiaoqiang Lu, and Xuelong Li. 2021. Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2017. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 863–871.

Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2018. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7405–7414.

Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.