MAR-Net: Motion-Assisted Reconstruction Network for Unsupervised Video Summarization

Yunzuo Zhang, *Member, IEEE*, Yameng Liu, Weili Kang, and Yuxin Zheng

***Abstract*—Video summarization targets to extract the most important segments from a video by spatiotemporal analysis. Previous methods primarily learn content within videos based on appearance information, with a rare discussion on the effective utilization of motion information, which is equally essential to video understanding. In this letter, we expound upon a Motion-Assisted Reconstruction Network (MAR-Net), which synergistically models appearance and motion information within videos for unsupervised video summarization without any manual annotations. MAR-Net notably comprises a Bidirectional Modality Encoder (BiME) and a Video Context Navigator (VCN). By integrating uni-modal and cross-modal feature aggregation into a unified module, BiME allows for exploring sophisticated dependency relationships among features through a bidirectional attention mechanism. VCN can promote the semantic consistency between the cross-modal contexts and the input video by a consistency loss term, alleviating the noisy impact within the motion stream. Empirical results conducted on benchmark datasets demonstrate that MAR-Net outperforms other state-of-the-art methods.**

***Index Terms*—Video summarization, motion information, attention mechanism, semantic consistency.**

I. Introduction

VIDEO summarization is a meaningful yet understudied topic in computer vision [1], which aims at automatically picking meaningful and informative segments from a video sequence, thus offering a time-efficient and engaging viewing experience. Along with numerous contributions proposed, video summarization increasingly serves a fundamental role in video browsing and retrieval [2].

Modern video summarization methods can be categorized into frame-level methods and shot-level methods. Frame-level methods usually first extract the visual feature of each frame, followed by popular aggregation tips (e.g., Recursive Neural Networks) to model frame-level contextual information. As an instance, Zhou et al. [3] aggregated long-range temporal cues by feeding frame-level visual features into Long Short-Term Memory. Liu et al. [4] targeted to directly encode spatiotemporal information within videos exploiting 3D convolution networks. Despite encouraging improvements made in summarizing videos, these methods overlook the visual similarity and continuity [5] across adjacent frames and are constrained in their capability to accurately assess the importance of each frame. Instead of predicting scores for all frames, shot-level summarization methods first segment the entire video sequence into several non-overlapping shots and further label them shot-level importance scores. For instance, Zhao et al. [6] adopted a reconstructive sequence-graph model to capture feature dependencies across shots. To promote informative shot-level representation learning, Zhang et al. [7] introduced a unified framework that jointly explores reinforcement and contrastive learning. However, these methods primarily achieve content understanding based on static appearance information, while omitting dynamic motion information which attracts humans for comprehending activities taking place within videos [8], [9].

Optical flow is usually leveraged to describe the motion information within videos [10], [11]. However, existing optical flow-based summarization methods mainly utilize simple strategies (e.g., two-stream network) to handle features of different modalities. This cannot allow for learning sophisticated dependency relationships across different modalities due to limited information interaction. Also, noisy information within optical flow caused by the dataset gap might further negatively affect the fine content understanding.

To address these issues, this letter presents a Motion-Assisted Reconstruction Network termed MAR-Net to effectively utilize both appearance and motion information. In a nutshell, we have the following contributions. (1) We present a novel deep learning-based pipeline termed MAR-Net. To the best of our knowledge, this is the first method that adopts the attention mechanism to achieve deep information exchange across appearance and motion for video summarization. (2) We design a Bidirectional Modality Encoder (BiME) that can effectively conduct feature aggregation within and between modalities, allowing for exploring sophisticated context dependencies. (3) We propose a Video Context Navigator (VCN) to promote the semantic consistency between the cross-modal contexts and the input video by a consistency loss term, reducing the negative impact within the motion stream.

Manuscript received 25 April 2023; revised 11 August 2023; accepted 4 September 2023. Date of publication 8 September 2023; date of current version 18 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 61702347, and 62027801, in part by the Natural Science Foundation of Hebei Province under Grants F2022210007, and F2017210161, in part by the Science and Technology Project of Hebei Education Department under Grants ZD2022100, and QN2017132, and in part by the Central Guidance on Local Science and Technology Development Fund under Grant 226Z0501G. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hantao Liu. *(Corresponding author: Yunzuo Zhang.)*

The authors are with the Shijiazhuang Tiedao University, Shijiazhuang 050043, China (e-mail: zhangyunzuo888@sina.com; liuym4647@sina.com; wayleek@sina.com; zhengyuxin7@sina.com).

Digital Object Identifier 10.1109/LSP.2023.3313091

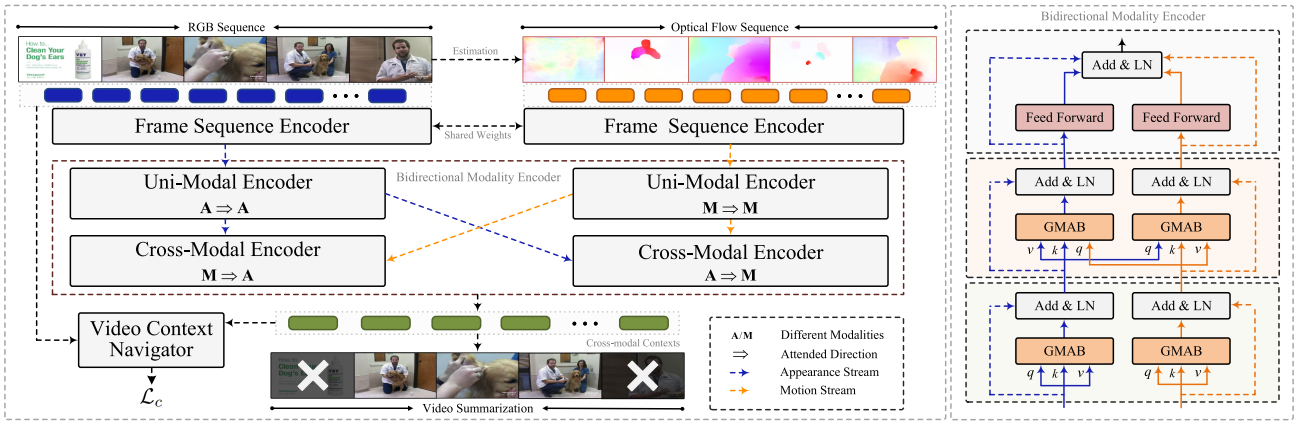


Fig. 1. Overall architecture of the proposed MAR-Net.

II. Proposed Approach

A. Overview

Fig. 1 presents the overall architecture of the proposed MAR-Net. Given an untrimmed video with T frames, we utilize a feature extractor to model appearance features $\boldsymbol{F}^A \in \mathbb{R}^{T\times d}$ and motion features $\boldsymbol{F}^M \in \mathbb{R}^{T\times d}$ from RGB and optical flow images, where d represents the dimension. The summary candidate $\boldsymbol{S} = \{x_i | a_{x_i} = 1, i = 1, \ldots, |\boldsymbol{S}|\}$ is selected by performing the Bernoulli sampling over predicted importance scores $\boldsymbol{P} = \{p_i\} \in \mathbb{R}^N$, which are generated using a prediction head formed by Linear $\rightarrow$ Sigmoid for cross-modal contexts. Here N denotes the number of non-overlapping shots detected by KTS [12]. $a_{x_i} \in \{0, 1\}$ indicates whether the x_i-th shot is selected into $\boldsymbol{S}$.

B. Frame Sequence Encoder

The Frame Sequence Encoder (FSE) is exploited to model local frames and aggregate them into shot-level latent representations, thereby enabling the exploitation of visual similarity and continuity between adjacent frames. Concretely, for feature sequence $\boldsymbol{F}^\tau$ ($\tau \in \{A, M\}$) with visual change points $\{(c_i^s, c_i^e)\}_{i=1}^N$, FSE feeds features within each shot into a Gated Recurrent Unit (GRU) to capture both forward and backward temporal cues. Here c_i^s and c_i^e refer to the starting and ending indices of i-th shot, respectively. Subsequently, the bidirectional hidden states are concatenated together to output shot-level representations. We define the process as follows:

$$\overrightarrow{\boldsymbol{h}_i^\tau} = \overrightarrow{GRU}(\boldsymbol{O}_i^\tau), \overleftarrow{\boldsymbol{h}_i^\tau} = \overleftarrow{GRU}(\boldsymbol{O}_i^\tau), \boldsymbol{h}_i^\tau = [\overrightarrow{\boldsymbol{h}_i^\tau} \| \overleftarrow{\boldsymbol{h}_i^\tau}] \quad (1)$$

where $\overrightarrow{\boldsymbol{h}_i^\tau}$ and $\overleftarrow{\boldsymbol{h}_i^\tau}$ are the forward and backward hidden features, respectively. $\boldsymbol{O}_i^\tau$ is the frame-level feature sequence corresponding to i-th shot. $\boldsymbol{h}_i^\tau \in \mathbb{R}^{1\times d}$ is the final latent representation after concatenation.

C. Bidirectional Modality Encoder

The pipeline of BiME is shown at the right of Fig. 1, which consists of two dominant structures including uni-modal encoders (UMEs) that contextualize features in each individual stream globally, and cross-modal encoders (CMEs) that facilitate the propagation of information across modalities. Technically, to retain position information, we begin by combining shot-level representations $\boldsymbol{H}^\tau = \{\boldsymbol{h}_i^\tau\}_{i=1}^N$ and position embeddings $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, forming position-sensitive representations $\boldsymbol{Z}^\tau$. Next, these representations are fed into the UMEs to learn self-attended feature dependencies $\boldsymbol{Z}_g^\tau \in \mathbb{R}^{N\times d}$ through a Generic Multi-Head Attention Block (GMAB) $\Gamma(\cdot)$, with the following procedure:

$$\boldsymbol{Z}_g^\tau = LN(\Gamma(\boldsymbol{Z}^\tau; \boldsymbol{Z}^\tau; \boldsymbol{Z}^\tau) + \boldsymbol{Z}^\tau) \quad (2)$$

where $\Gamma(\cdot)$ computes refined representations $\boldsymbol{Y}$ regarding query (q), key (k), and value (v), respectively, given by the following equations:

$$\boldsymbol{Y}_i = softmax\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^T}{\sqrt{d}}\right)\boldsymbol{V}_i \quad (3)$$

$$\boldsymbol{Y} = concat(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_r)\boldsymbol{W}^o \quad (4)$$

where $\boldsymbol{Y}_i$ is the result in i-th head. $\boldsymbol{Q}_i$, $\boldsymbol{K}_i$, and $\boldsymbol{V}_i$ are mapped matrices through linear transforms from the query, key, and value. r is the number of attention heads and $\boldsymbol{W}^o$ indicates learnable weights.

Subsequently, CMEs are utilized to jointly model appearance and motion information by exchanging queries of two streams. This encoder adopts a bidirectional modeling structure, allowing static appearance information to attend to dynamic motion information represented as $A \Rightarrow M$ and vice versa represented as $M \Rightarrow A$. By doing so, more effective and informative representations are obtained. Formally, the cross-attended features $\boldsymbol{U}_{A\Rightarrow M} \in \mathbb{R}^{N\times d}$ and $\boldsymbol{U}_{M\Rightarrow A} \in \mathbb{R}^{N\times d}$ can be defined as follows:

$$\boldsymbol{U}_{A\Rightarrow M} = LN(\Gamma(\boldsymbol{Z}_g^A; \boldsymbol{Z}_g^M; \boldsymbol{Z}_g^M) + \boldsymbol{Z}_g^M) \quad (5)$$

$$\boldsymbol{U}_{M\Rightarrow A} = LN(\Gamma(\boldsymbol{Z}_g^M; \boldsymbol{Z}_g^A; \boldsymbol{Z}_g^A) + \boldsymbol{Z}_g^A) \quad (6)$$

Eventually, feed-forward networks $FFN(\cdot)$ are employed to convey deep features followed by residual connections and a summation operation, outputting cross-modal contexts $\boldsymbol{U}_{final} \in \mathbb{R}^{N\times d}$ that covers both static and dynamic information within

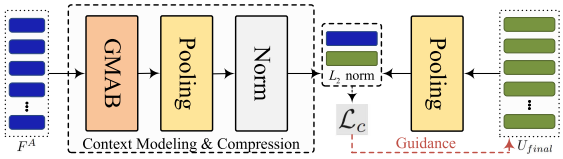


Fig. 2. Illustration of the VCN architecture.

videos:

$$\boldsymbol{U}_{final} = LN(FFN(\boldsymbol{U}_{A\Rightarrow M}) + \boldsymbol{U}_{A\Rightarrow M} + FFN(\boldsymbol{U}_{M\Rightarrow A}) + \boldsymbol{U}_{M\Rightarrow A}) \quad (7)$$

D. Video Context Navigator

Optical flow models can effectively capture motion information by rendering continuous RGB images. However, due to the complexity of realistic scenarios as well as the domain gap across datasets [13], much noise is inevitably introduced in the motion stream caused by inaccurate estimation, leading to semantic inconsistency across aggregated features and input videos.

In fact, activities occurring within videos can be easily inferred from the appearance stream. This motivates us to devise VCN, shown in Fig. 2, to alleviate the above issue. In particular, the frame-level appearance features $\boldsymbol{F}^A$ are aggregated into global contextual information. Considering the long range of sequential data, we employ GMAB followed by global average pooling along the temporal dimension and layernorm to form the appearance context vector $\boldsymbol{U}^v_{sum} \in \mathbb{R}^d$. Next, the same pooling operation is performed on $\boldsymbol{U}_{final}$ to enable loss calculation, outputting $\boldsymbol{U}^v_{final} \in \mathbb{R}^d$. With the two components, we define the consistency loss $\mathcal{L}_c$ as:

$$\mathcal{L}_c = \|\boldsymbol{U}^v_{final} - \boldsymbol{U}^v_{sum}\|_2 \quad (8)$$

where $\|\cdot\|_2$ indicate L_2 norm. Leveraging this module, our method can alleviate the negative influence within the motion stream by facilitating semantic consistency across cross-modal dependencies and pure appearance dependencies.

E. Network Training

The objective mainly includes three loss terms including reinforcement reward $\mathcal{R}$, regularization term $\mathcal{L}_s$, and consistency loss $\mathcal{L}_c$. To verify improvement, we adopt diversity and representativeness rewards used in [7] according to shot-level representations $\boldsymbol{H}^A$ to select semantically representative and diverse shots. The regularization term is computed as $\mathcal{L}_s = \|\frac{1}{N}\sum_{i=1}^{N} p_i - 0.5\|^2$ to control the proportion of selected shots. Formally, the overall loss $\mathcal{L}$ can be defined below:

$$\mathcal{L} = \alpha\mathcal{L}_c + \mathcal{L}_s - \mathcal{R}\,(+\mathcal{L}_m) \quad (9)$$

where $\mathcal{L}_m$ denotes the MSE loss of predicted scores against human-created scores when MAR-Net is extended to a supervised learning model. α is a hyper-parameter and empirically set to 0.01 in this letter.

TABLE I
F-SCORE COMPARISON WITH DIFFERENT UNSUPERVISED METHODS

Method	Venue	Modality	SumMe	TVSum
GAN$_{dpp}$[19]	CVPR'17	RGB	0.391	0.517
DR-DSN [3]	AAAI'18	RGB	0.414	0.576
Cycle-SUM [20]	AAAI'19	RGB	0.419	0.576
ACCGN [21]	ACM MM'19	RGB	0.460	0.585
SUM-GDA [22]	PR'21	RGB+Flow	0.494	0.602
RSGN [6]	TPAMI'22	RGB	0.423	0.580
3DST-UNet [4]	TIP'22	RGB	0.446	0.581
RCL [7]	SPL'22	RGB	0.486	0.584
MAR-Net	Ours	RGB+Flow	0.497	0.604

Red, blue, and green represent the best three records.

III. EXPERIMENTS

A. Datasets and Metrics

To showcase the effectiveness of the proposed method, we conduct experiments on the commonly used SumMe [14] and TVSum [15] datasets. We follow previous methods [6], adopting F-score and rank-order statistics [16] as our evaluation. F-score evaluation can measure the overlap between the generated summary and the human summary, which are defined as F-score $= (2 \times P \times R)/(P + R)$, where P and R are the precision and recall of generated summary against the human summary. Rank-order statistics are used to measure the correlation between predicted scores and annotated scores, overcoming the limitations of F-score evaluation.

B. Implementation Details

Consistent with previous methods [3], [7], we partition each dataset into two disjoint parts: 80% of videos for training and the rest for testing. We report the average F-score using standard 5-fold cross-validation to ensure fairness. The optical flow is rendered by the state-of-the-art method [17]. We adopt pre-trained GoogLeNet [18] to extract features and d is set to 1024. The number of attention heads r is set to 8. The dropout rate is set to 0.1. We optimize our model by Adam optimizer with a learning rate of 5×10^{-5}.

C. Comparisons With State-of-The-Arts

Table I presents the experimental results of our unsupervised model against other state-of-the-art methods. MAR-Net achieves the best summarization performance on the SumMe and TVSum datasets. RCL [7] is a state-of-the-art shot-level method, the same as our method. However, under the strictly fair comparison setting, MAR-Net significantly surpasses it on two datasets, which can be attributed to the fact that combining appearance and motion information together can further benefit the capability of video understanding. Also, our method outperforms SUM-GDA [22] based on RGB and optical flow modalities since our elaborate architecture allows for exploring sophisticated dependencies across modalities. Table II reports the results by easily extending MAR-Net to a supervised method. It can be seen that our method still

TABLE II
F-SCORE COMPARISON WITH DIFFERENT SUPERVISED METHODS

Method	Venue	Modality	SumMe	TVSum
GAN$_{sup}$ [19]	CVPR'17	RGB	0.417	0.563
DR-DSN$_{sup}$ [3]	AAAI'18	RGB	0.421	0.581
ACCGN$_{sup}$ [21]	ACM MM'19	RGB	0.472	0.594
SUM-GDA$_{sup}$[22]	PR'21	RGB+Flow	0.482	0.610
RSGN$_{sup}$ [6]	TPAMI'22	RGB	0.450	0.601
3DST-UNet$_{sup}$ [4]	TIP'22	RGB	0.474	0.583
RCL$_{sup}$ [7]	SPL'22	RGB	0.486	0.600
HTM [23]	Neurocom.'22	RGB	0.441	0.601
LHMA [24]	PR'22	RGB+Flow	0.514	0.615
MAR-Net$_{sup}$	Ours	RGB+Flow	0.511	0.610

Red, blue, and green represent the best three records.

TABLE III
RANK-ORDER STATISTICS COMPARISON WITH DIFFERENT METHODS

Method	Paradigm	Kendall's τ	Spearman's ρ
Random [16]	-	0.000	0.000
Human [16]	-	0.177	0.204
dppLSTM [25]	Supervised	0.042	0.055
DR-DSN [3]	Unsupervised	0.020	0.026
RSGN [6]	Unsupervised	0.048	0.052
RSGN$_{sup}$[6]	Supervised	0.083	0.090
HTM [23]	Supervised	0.096	0.107
MAR-Net	Unsupervised	0.040	0.055
MAR-Net$_{sup}$	Supervised	**0.154**	**0.215**

TABLE IV
ABLATION STUDY ON FEATURE STREAM

Exp.	Feature Stream		SumMe	TVSum
	Appearance	Motion		
1	✓		0.487	0.593
2		✓	0.485	0.597
3	✓	✓	**0.497**	**0.604**

exhibits comparable performance, demonstrating the excellent capability of learning guidance information from human annotations.

Furthermore, the rank-order statistics of different methods are reported in Table III. Our unsupervised model achieves promising results in both Kendall's τ and Spearman's ρ. Interestingly, the correlation coefficients achieved by the supervised model are considerably higher than those of the unsupervised model, even comparable to the human summary. This further confirms the effectiveness of learning human-created guidance.

D. Ablation Study

1) Study on Feature Stream: We investigate the impact of appearance and motion features by inputting only one modality into the network. The results are shown in Table IV. It is evident that our ultimate architecture (Exp. 3) yields the best performance, illustrating the beneficial effect of either appearance or motion information in videos.

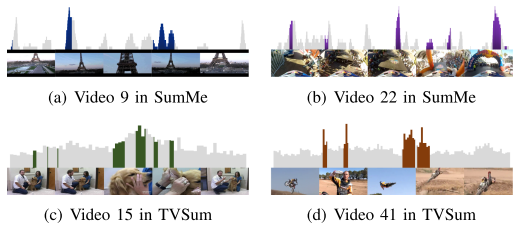
2) Study on Attended Direction: Table V shows the ablation results of attended direction in BiME. According to reported values, unidirectional attention can degrade summarization performance on the SumMe and TVSum datasets, respectively. Conversely, bidirectional interaction allows for more comprehensive feature aggregation across different modalities.

TABLE V
ABLATION STUDY ON ATTENDED DIRECTION

Exp.	Attended Direction		SumMe	TVSum
	A⇒ M	M⇐ A		
1	✓		0.471	0.597
2		✓	0.470	0.600
3	✓	✓	**0.497**	**0.604**

TABLE VI
ABLATION STUDY ON CONSISTENCY LOSS

Exp.	Method	Unsupervised		Supervised	
		SumMe	TVSum	SumMe	TVSum
1	MAR-Net w/o $\mathcal{L}_c$	0.467	0.596	0.497	0.600
2	+ $\mathcal{L}'_c$	0.486	0.602	0.505	0.608
3	+ $\mathcal{L}_c$	**0.497**	**0.604**	**0.511**	**0.610**



(a) Video 9 in SumMe (b) Video 22 in SumMe

(c) Video 15 in TVSum (d) Video 41 in TVSum

Fig. 3. Example visualization results generated by our method.

3) Study on Consistency Loss: Table VI investigates the impact of the consistency loss of both unsupervised and supervised models. Here $\mathcal{L}'_c$ is computed without modeling temporal cues within appearance features, similar to [26]. It can be seen that the model (Exp. 3) combining $\mathcal{L}_c$ significantly performs better than that without consistency loss (Exp. 1), which can be attributed to the effective guidance of VCN. Moreover, the performance improvement between Exp. 2 and Exp. 3 highlights the suitability of guiding cross-modal interaction features through temporal cues in visual features.

E. Qualitative Results

The visualization results are depicted in Fig. 3, where the manual annotations and summaries generated by MAR-Net are represented by light gray and colored bars, respectively. As evidenced by the visualization, the proposed method is capable of identifying and extracting important segments from videos, from which we can easily infer activities taking place in these videos.

IV. CONCLUSION

This letter has presented a Motion-Assisted Reconstruction Network termed MAR-Net to tackle unsupervised video summarization. We first propose a Bidirectional Modality Encoder (BiME), an architecture that integrates uni-modal and cross-modal interactions, to effectively achieve video understanding according to appearance and motion information through a bidirectional attention mechanism. Furthermore, a Video Context Navigator (VCN) is devised to promote semantic consistency between cross-modal contexts and the input video by a developed consistency loss term, further boosting summarization performance. Extensive experiments conducted on benchmarks validate the superiority of the proposed method.

REFERENCES

[1] Y. Zhang, Z. Song, and W. Li, "Enhancement multi-module network for few-shot leaky cable fixture detection in railway tunnel," *Signal Process. Image Commun.*, vol. 113, 2023, Art. no. 116943.

[2] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder–decoder networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1709–1717, Jun. 2020.

[3] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[4] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, and B. Kainz, "Video summarization through reinforcement learning with a 3D spatio-temporal U-Net," *IEEE Trans. Image Process.*, vol. 31, pp. 1573–1586, 2022.

[5] H. Fu and H. Wang, "Self-attention binary neural tree for video summarization," *Pattern Recognit. Lett.*, vol. 143, pp. 19–26, 2021.

[6] B. Zhao, H. Li, X. Lu, and X. Li, "Reconstructive sequence-graph network for video summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2793–2801, May 2022.

[7] Y. Zhang, Y. Liu, P. Zhu, and W. Kang, "Joint reinforcement and contrastive learning for unsupervised video summarization," *IEEE Signal Process. Lett.*, vol. 29, pp. 2587–2591, 2022.

[8] Y. Zhang, R. Tao, and Y. Wang, "Motion-state-adaptive video summarization via spatiotemporal analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1340–1352, Jun. 2017.

[9] E. G. Sartinas, E. Z. Psarakis, and D. I. Kosmopoulos, "Motion-based sign language video summarization using curvature and torsion," 2023, *arXiv:2305.16801*.

[10] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7274–7283.

[11] D. Min, C. Zhang, Y. Lu, K. Fu, and Q. Zhao, "Mutual-guidance transformer-embedding network for video salient object detection," *IEEE Signal Process. Lett.*, vol. 29, pp. 1674–1678, 2022.

[12] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 540–555.

[13] X. Gu, H. Chang, B. Ma, and S. Shan, "Motion feature aggregation for video-based person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 3908–3919, 2022.

[14] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 505–520.

[15] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: Summarizing web videos using titles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5179–5187.

[16] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7596–7604.

[17] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 402–419.

[18] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[19] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 202–211.

[20] L. Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, "Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9143–9150.

[21] X. He et al., "Unsupervised video summarization with attentive conditional generative adversarial networks," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2296–2304.

[22] P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, "Exploring global diverse attention via pairwise temporal relation for video summarization," *Pattern Recognit.*, vol. 111, 2021, Art. no. 107677.

[23] B. Zhao, M. Gong, and X. Li, "Hierarchical multimodal transformer to summarize videos," *Neurocomputing*, vol. 468, pp. 360–369, 2022.

[24] W. Zhu, J. Lu, Y. Han, and J. Zhou, "Learning multiscale hierarchical attention for video summarization," *Pattern Recognit.*, vol. 122, 2022, Art. no. 108312.

[25] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 766–782.

[26] A. Phaphuangwittayakul, Y. Guo, F. Ying, W. Xu, and Z. Zheng, "Self-attention recurrent summarization network with reinforcement learning for video summarization task," in *Proc. Int. Conf. Multimedia Expo.*, 2021, pp. 1–6.