Highlights

Domain Adaptive Video Summarization using Generalized Transformer

Ziyi Wang[a], Yubo Zhu[a], Xinxiao Wu[a,b]

- We propose a two-stage domain adaptation framework for cross-domain video summarization, addressing domain shift by enhancing the model's generalization and adaptation abilities.
- At the vendor stage, we introduce a straightforward yet effective regularized feature encoder based on Transformer to improve the model's generalization ability across diverse domains.
- At the client stage, we design a discrepancy reduction loss with confidence weighting to mitigate domain shift by adapting the model to a specific target domain.
- Experiments across various datasets and evaluation metrics demonstrate that our method outperforms the state-of-the-art methods.

Domain Adaptive Video Summarization using Generalized Transformer

Ziyi Wang[a], Yubo Zhu[a], Xinxiao Wu[a,b]

[a]*Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China*
[b]*Guangdong Provincial Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen, China*

Abstract

Video summarization aims to extract salient segments from videos to construct concise and comprehensive synopses. Despite significant advancements, the diversity of video content and the constraint of limited training data pose challenges when applying trained models to new scenarios, often resulting in the domain shift problem. To address this challenge, we propose a domain adaptation framework tailored to video summarization from two aspects: (a) enhancing the generalization ability and (b) improving the adaptive ability of video summarization models. Specifically, we design a simple yet effective regularized feature encoder based on Transformer, where an averaging operation on attention weights serves as a form of regularization. This method mitigates overfitting to domain-specific cues and encourages the learning of more generalizable representations across diverse domains. Furthermore, we introduce a novel discrepancy reduction loss that aligns the distribution of inter-frame feature similarities and inter-frame prediction similarities, combined with a confidence weighting strategy, to adapt the regularized encoder to target domains and mitigate domain shift. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness of our method. Our method achieves state-of-the-art performance under various settings on TVSum and SumMe, and obtains the best results on the transfer setting of Mr.HiSum.

Keywords:
Video summarization; Domain adaptation; Transformer

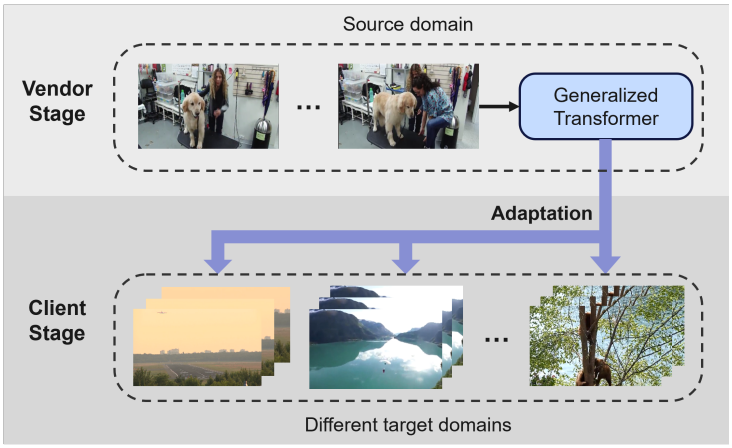


Figure 1: The domain adaptation framework for cross-domain video summarization. The vendor stage improves the generalization of the model in different domains, and the client stage alleviates the domain shift by adapting the model to a specific target domain.

1. Introduction

Video summarization focuses on automatically creating a concise and comprehensive synopsis of an input video by extracting its significant and representative segments. This field finds extensive applications in video-sharing platforms and the video surveillance industry. Many existing supervised video summarization methods [1, 2, 3] assume that the training and testing data follow the same distribution. However, this assumption is impractical in real-world scenarios, since the target testing domain often presents a distinct data distribution from the source training domain. Such differences, characterized by variations in appearance, illumination, and background, are commonly referred to as domain shift.

To address the domain shift problem, cross-domain methods have emerged as a solution by transferring knowledge from a labeled source domain to an unlabeled target domain. Zhang *et al.* [4] borrow the ideas from object recognition [5, 6, 7], and introduce a simple cross-domain method [8] called CORAL into video summarization for the first time, which minimizes domain shift by aligning the second-order statistics of source and target distributions, without requiring any target labels. In addition, Ho *et al.* [9], inspired by [10], explore the cross-domain feature embedding. They propose a novel deep neural network architecture for describing and discriminating vital spatiotemporal information across videos with different points of view.

While these methods achieve promising results, they only consider improving the adaptive ability of the video summarization model. In contrast, our method tackles the domain shift problem from two aspects: enhancing the generalization ability and improving the adaptive ability, achieving better performance.

This paper presents an innovative domain adaptation framework designed for cross-domain video summarization, which tackles the intricate problem of domain shift by focusing on two pivotal aspects: enhancing the generalization ability and improving the adaptive ability. We describe this framework through two stages: the vendor stage and the client stage, as shown in Figure 1. The vendor stage is dedicated to augmenting the generalization ability of the source model across diverse domains, and the client stage focuses on adapting the source model to a specific target domain. In the vendor stage, the source model is trained to generalize well across multiple domains without accessing target data. In contrast, the client stage aims at adapting the source model to a specific target domain using only unlabeled target data. To achieve effective domain generalization and adaptation, both stages incorporate tailored learning strategies.

In the vendor stage, we propose a simple yet effective regularized feature encoder based on Transformer [11] to learn robust frame features across diverse domains. Traditional Transformer-based attention mechanisms compute pairwise similarities between frames. While this formulation is theoretically capable of modeling global information, in practice, especially when faced with domain shifts, the attention may become biased toward pairwise visual similarities, potentially leading to overfitting to domain-specific cues. To address this, we design an average pooling operation on the attention weights, replacing the standard value-weighted summation in self-attention. The averaging operation acts as a form of regularization, mitigating the risk of the model overfitting to spurious pairwise similarities and encouraging the learning of more generalizable representations. By smoothing the attention distribution, this regularization enhances the model's generalization ability across diverse domains.

The effectiveness of the proposed method is demonstrated through comprehensive experiments on four benchmark datasets: TVSum [12], SumMe [13], FPVSum [9], and Mr.HiSum [14]. Our method consistently achieves competitive results when compared with state-of-the-art methods. The main contributions of our work are summarized as follows:

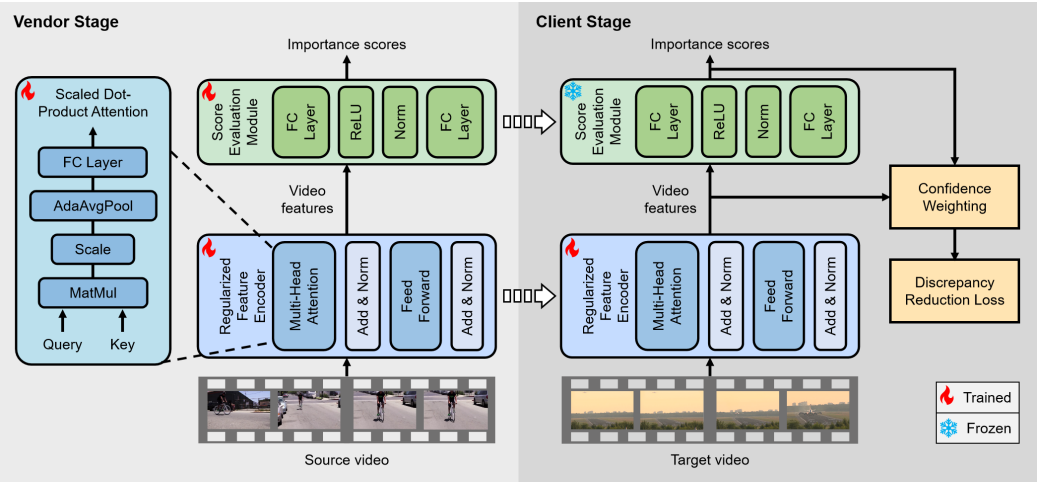


Figure 2: The framework of our method. At the vendor stage, we propose a straightforward yet effective regularized feature encoder based on Transformer to extract video features. At the client stage, we design a discrepancy reduction loss to adapt the regularized feature encoder to target domains and a confidence weighting strategy to assign different weights to samples during the computation of the discrepancy reduction loss.

- We first propose a two-stage domain adaptation framework for cross-domain video summarization, addressing domain shift by enhancing the model's generalization and adaptation abilities.
- At the vendor stage, we introduce a straightforward yet effective regularized feature encoder based on Transformer to improve generalization across diverse domains.
- At the client stage, we design a discrepancy reduction loss with confidence weighting to mitigate domain shift by adapting the model to a specific target domain.

The remainder of this paper is organized as follows. Section 2 reviews the related work on video summarization and cross-domain methods in video summarization. Section 3 details the proposed method, including the vendor stage and the client stage. Section 4 presents extensive experiments to evaluate the effectiveness of the proposed method, and Section 5 points out the limitations and future work of this paper. Finally, Section 6 concludes the paper.

2. Related Work

2.1. Video Summarization

Early video summarization methods generate video summaries by clustering low-level visual features, such as color histograms [15], spatio-temporal features [16], and motion cues [17]. With the considerable progress of deep learning in video processing and understanding, deep learning-based methods have been proposed. Zhao *et al.* [18] propose a fixed-length hierarchical RNN, while Zhao *et al.* [19] introduce a hierarchical structure-adaptive LSTM to model the underlying hierarchical structure of videos. More recently, the self-attention mechanism has been widely employed in video summarization. Zhong *et al.* [20] introduce a self-attention mechanism to address the difficulty of modeling temporal information over long time spans. Zhu *et al.* [3] propose a multiscale hierarchical attention approach to learn local and global information. Beyond general video summarization methods, recent studies incorporate domain knowledge, such as audio-visual cues and user preferences, to enable personalized and dynamic summarization [21, 22, 23, 24].

All these methods heavily rely on the assumption that the training and testing data follow the same distribution. To transcend this assumption and render video summarization more practical for real-world applications, we propose a domain adaptation framework tailored to cross-domain video summarization, which aims to alleviate the domain shift from the aspects of generalization ability and adaptive ability.

2.2. Cross-domain technology in Video Summarization

Cross-domain techniques have gained increasing attention in video summarization, enabling knowledge transfer from the source domain to the target domain. Borrowing the ideas from object recognition [5, 6, 7], Zhang *et al.* [4] first introduce a simple cross-domain method [8] called CORAL into video summarization to reduce the data distribution discrepancy among different datasets. More concretely, they align the distributions by re-coloring whitened source features with the covariance of the target distribution. In addition, inspired by [10], Ho *et al.* [9] propose a novel deep neural network architecture for describing and discriminating vital spatiotemporal information across videos with different points of view. More specifically, they perform cross-domain feature embedding and transfer representative highlight information across different domains through an auxiliary reconstruction task. Additionally, Jiang and Mu [25] utilize external training data from the video

moment localization task to alleviate the lack of labeled data in video summarization and improve the generalization ability of their model across different datasets.

While these methods achieve promising results, they only consider improving the adaptive ability of the video summarization model. In contrast, our method tackles the domain shift problem from two aspects: enhancing the generalization ability and improving the adaptive ability, achieving better performance.

3. Our Method

3.1. Overview

Video summarization aims to extract the most significant and representative segments of a video to create a concise summary. The input consists of a sequence of video frames, and the goal is to predict an importance score for each frame that indicates its relevance within the context of the entire video. These scores are then used to calculate the average importance of each pre-divided segment of fixed length in the video, to select key segments to make up the video summary. In this paper, we propose a two-stage method, consisting of a vendor stage and a client stage, to enhance both the generalization and adaptation capabilities of a video summarization model. The vendor stage focuses on learning a generalized Transformer model that improves the robustness of frame representations across different domains by introducing a regularization effect in the attention mechanism. The client stage aims to improve the model's performance on target-specific data by minimizing domain shifts and refining the model with unlabeled target video data. Figure 2 illustrates the overview of our method. Following prior works, we directly use frame-level features extracted from GoogLeNet [26] as the input to our model.

3.2. Vendor Stage

The vendor stage aims to enhance the generalization ability of the source model, preparing it for effective adaptation to the target domain. To achieve this, we focus on learning a generalized Transformer in which an averaging operation is applied to the attention weights. This averaging acts as a form of regularization, smoothing the attention distribution and reducing the model's reliance on potentially spurious pairwise similarities that are sensitive to domain-specific variations. By regularizing the attention mechanism in this

way, the model is encouraged to learn more generalizable and robust frame representations, which are less likely to overfit to the source domain and thus better suited for transfer to new domains. This regularization strategy is particularly beneficial in video summarization, where the diversity of content and limited training data can otherwise lead to poor generalization.

Building on this insight, we modify the self-attention operation in a single-layer Transformer to extract frame features by replacing the standard value-weighted summation with global aggregation of attention weights, which are computed from the dot product of the key and query vectors. Specifically, the proposed model consists of two modules: a regularized feature encoder and a score evaluation module. We choose a single-layer Transformer encoder as the regularized feature encoder, and the input sequence of frame features extracted by GoogLeNet [26] is denoted by $X = [v_1, v_2, ..., v_n]^\top$, $X \in \mathbb{R}^{n \times d_m}$, where n is the number of frames and d_m is the dimension of the frame feature. Then for the h-th head in the multi-head self-attention, we can obtain an attention weight matrix $A_h \in \mathbb{R}^{n \times n}$ by

$$Q_h = XW_h^Q,\ K_h = XW_h^K,\ h = 1, \cdots, H, \tag{1}$$

$$A_h = \text{Softmax}(\frac{Q_h K_h^\top}{\sqrt{d_k}}), \tag{2}$$

where H is the number of heads in the multi-head attention mechanism and $Q_h, K_h \in \mathbb{R}^{n \times d_k}$ denote the query and key matrices of the h-th attention head, $W_h^Q \in \mathbb{R}^{d_m \times d_k}$, $W_h^K \in \mathbb{R}^{d_m \times d_k}$ are the learnable attention parameters, and $d_k = d_m/H$.

To obtain the regularized attention of each frame, we use an adaptive average pooling layer to aggregate the similarity between each frame and other frames globally. Then, a fully connected layer is adopted to remap the regularized attention of all frames back to the same dimension as the query and key vectors, allowing the model to reallocate feature importance based on the pooled results. The process can be formulated as:

$$\hat{A}_h = \text{Pool}(A_h), \tag{3}$$

$$O_h = FC(\hat{A}_h), \tag{4}$$

where $\hat{A}_h \in \mathbb{R}^{n \times 1}$ denotes the regularized attention of all frames, $\text{Pool}(\cdot)$ denotes the adaptive average pooling layer, and $FC(\cdot)$ is a fully connected layer, and $O_h \in \mathbb{R}^{n \times d_k}$ denotes the result of the h-th attention head. In this way,

we reduce extreme values in the attention matrix, prevent individual frames from obtaining excessively high or low weights, and enhance the stability of the model.

After calculating the result of each self-attention head, we concatenate the attention result of each head and obtain the multi-head attention result by $M = Concat(O_1, O_2, ..., O_H)$, where $M \in \mathbb{R}^{n \times d_m}$. The multi-head attention result M is then fed into a non-linear three-layer network, which can be formulated as:

$$F = \max(0, MW_1 + b_1) \cdot W_2 + b_2, \tag{5}$$

where $F \in \mathbb{R}^{n \times d_m}$ denotes the frame features finally extracted by the proposed regularized feature encoder and $W_1 \in \mathbb{R}^{d_m \times d_f}$, $W_2 \in \mathbb{R}^{d_f \times d_m}$ are the learnable parameters for the non-linear network.

It is worth knowing that we apply residual connection when calculating the result of multi-head attention and the final feature, to preserve original frame-level information to a certain degree. Additionally, we attach Layer Normalization after each residual connection for better convergence. Residual connections and layer normalization are together represented as "Add & Norm" blocks in Figure 2, which are attached after the multi-head attention and the non-linear feed-forward layer.

Finally, the score evaluation module consists of two fully connected layers. The detailed formula is as follows:

$$\hat{S} = FC_2(\text{LayerNorm}(FC_1(F))) \tag{6}$$

where $FC_1(\cdot)$ is a fully connected layer with ReLU activation function to transform the dimension of F from $\mathbb{R}^{n \times d_m}$ to $\mathbb{R}^{n \times 128}$, LayerNorm$(\cdot)$ denotes layer normalization and $FC_2(\cdot)$ is a single linear layer to get the predicted importance score $\hat{S} \in \mathbb{R}^n$.

In training, we make use of the ground truth frame-level importance scores to supervise the learning process. Given the ground truth frame scores $S_{\text{gt}} \in \mathbb{R}^n$ and the predicted frame scores $\hat{S}$, we adopt the mean squared error (MSE) as the loss function:

$$\mathcal{L}_{\text{mse}} = \frac{1}{n} \left\| \hat{S} - S_{\text{gt}} \right\|_2^2. \tag{7}$$

3.3. Client Stage

The client stage serves the purpose of improving the model's adaptive ability, leveraging the potential of unlabeled target data. We design a novel

discrepancy reduction loss to adapt the regularized feature encoder to alleviate the domain shift and introduce a confidence weighting strategy that assigns different weights to samples during the computation of the discrepancy reduction loss.

3.3.1. Discrepancy Reduction Loss

We adapt the regularized feature encoder while keeping the score evaluation module unchanged to facilitate adaptation to the target domain. Specifically, we aim to better align the output features of the regularized feature encoder to the source feature distribution, thereby reducing the impact of domain shift and improving the accuracy of frame importance score predictions in the target domain. We find that the similarity between frame features is strongly correlated with the similarity between their importance scores in a well-trained source model for video summarization. Specifically, frame features that are close to each other in the feature space tend to have similar importance scores. Motivated by this, we propose a discrepancy reduction loss function to adapt the regularized feature encoder by constraining the pairwise similarities of target features to better align with the importance score similarities predicted by the source model.

We model the feature distribution of video frames as a pairwise similarity-based probability distribution P based on feature distances. Given the frame features $F = \{f_1, f_2, ..., f_n\}$ of the m-th video, their probability distributions are calculated as follows:

$$P_m(j|i) = \frac{\exp(D(f_i, f_j))}{\sum_{k \neq i} \exp(D(f_i, f_k))}, \ j \neq i, \ i = 1, ..., n, \tag{8}$$

where $D(f_i, f_j)$ is the distance metric between f_i and f_j, and we employ cosine similarity as the distance metric in our implementation. Similarly, we model the importance score distribution as another probability distribution P' based on the distances between predicted importance scores. Given the predicted importance scores $\hat{S} = \{\hat{s}_1, \hat{s}_2, ..., \hat{s}_n\}$ of the m-th video, their probability distributions are calculated by

$$P'_m(j|i) = \frac{\exp(D'(\hat{s}_i, \hat{s}_j))}{\sum_{k \neq i} \exp(D'(\hat{s}_i, \hat{s}_k))}, j \neq i, i = 1, ..., n, \tag{9}$$

where $D'(\hat{s}_i, \hat{s}_j) = -|\hat{s}_i - \hat{s}_j|$ is the distance metric for importance scores. A higher similarity between importance scores results in a greater distance,

ensuring that the probability distributions reflect their relative importance differences.

Finally, we use KL divergence to measure the discrepancy between the feature distribution and the importance score distribution, i.e., the discrepancy reduction loss $\mathcal{L}_d$ of the m-th video, formulated as

$$\mathcal{L}_d = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} P'_m(j|i)\log(\frac{P'_m(j|i)}{P_m(j|i)}). \tag{10}$$

We minimize the discrepancy reduction loss to constrain the similarities of target features between frames to be closer to those of importance scores, as predicted by the source model for the same frame pairs.

3.3.2. Confidence Weighting Strategy

To further improve performance, we design a confidence weighting strategy that assigns different weights to samples according to their confidence during the computation of the discrepancy reduction loss. Specifically, we use the KL divergence between the target feature distribution and the importance score distribution predicted by the source model to assess the confidence of each video frame. The smaller the KL divergence, the higher the confidence of the sample. Next, we use the Softmax function to calculate the weight of each video frame sample. For the m-th target video with n frames, the confidence weight is calculated by

$$\begin{aligned} Conf_i &= -\sum_{j\neq i} P'_m(j|i)\log(\frac{P'_m(j|i)}{P_m(j|i)}), i = 1, ..., n, \\ W_i &= \frac{\exp(Conf_i)}{\sum_j \exp(Conf_j)}, i = 1, ..., n, \\ W'_i &= \frac{\exp(-Conf_i)}{\sum_j \exp(-Conf_j)}, i = 1, ..., n, \end{aligned} \tag{11}$$

where W represents a positive weight, increasing with higher confidence. W' is a negative weight, which increases as confidence decreases.

We incorporate these confidence weights into the discrepancy reduction loss from two perspectives. When measuring the discrepancy between the feature distribution and the importance score distribution, the positive weight W_i helps refine the discrepancy measurement, making it more reliable. When calculating the total loss, the negative weight W'_i emphasizes low-confidence

Table 1: Different settings on the TVSum, SumMe and FPVSum datasets.

Dataset	Setting	Training Set	Testing Set
SumMe	Canonical	80% SumMe	20% SumMe
	Augmented	80% SumMe + TVSum + OVP + YouTube	20% SumMe
	Transfer	TVSum + OVP + YouTube	SumMe
TVSum	Canonical	80% TVSum	20% TVSum
	Augmented	80% TVSum + SumMe + OVP + YouTube	20% TVSum
	Transfer	SumMe + OVP + YouTube	TVSum
FPVSum	–	25% FPVSum + TVSum + 80% SumMe	20% SumMe

samples, ensuring they contribute more to the optimization. The weighted discrepancy reduction loss for the m-th target video is given by

$$\mathcal{L}_d^w = n \sum_{i=1}^{n} W_i' \sum_{j \neq i} W_j P_m'(j|i) \log\left(\frac{P_m'(j|i)}{P_m(j|i)}\right). \tag{12}$$

3.4. Video Summary Generation

After the vendor stage and client stage training processes, the learned model is used to perform video summarization in a unified pipeline. Specifically, given an input video, we use GoogLeNet to extract frame features as the pre-processing stage. Then the regularized feature encoder predicts an importance score for each frame, reflecting its relevance to the overall video content. These importance scores are then averaged over pre-divided segments, and the segments with the highest average importance scores are selected to form the final video summary. This two-stage design ensures that the model can first capture generalized representations across domains and subsequently adapt to the target domain's distribution, leading to more accurate and robust summarization results.

4. Experiment

4.1. Dataset

Our method's effectiveness is demonstrated through comprehensive experiments on four benchmark datasets: TVSum [12], SumMe [13], FPVSum [9], and Mr.HiSum [14]. TVSum contains 50 videos sourced from

YouTube, each annotated with shot-level importance scores by 20 users. These videos cover 10 distinct categories, with 5 videos per category. The categories include News, How-to/tutorials, Documentaries, Vlogs, Egocentric videos, User-generated content, Flash mob gatherings, Educational content, Interviews, and Event logs. SumMe contains 25 user-generated videos, ranging from 1 to 6.5 minutes in length. These videos cover a variety of topics, including Holidays, Events, Sports, Airplane landings, and Extreme sports. Following the protocol in [4], we introduce YouTube [15] and Open Video Project (OVP) [15, 27] datasets to build the FPVSum dataset, which comprises 56 labeled videos.

The details are shown in Table 1, where the transfer setting divides completely different datasets into training and testing sets to simulate cross-domain scenarios. For FPVSum, we introduce third-person videos from TVSum and SumMe into the training set, and first-person videos into the testing set, following the protocol in [9], where domain shifts are caused by different perspectives. The transfer setting of both SumMe and TVSum, together with the FPVSum dataset, is used as the domain adaptation setting.

In addition, we conduct experiments on Mr.HiSum [14], a large-scale video summarization dataset. Mr.HiSum consists of 31,892 videos sourced from YouTube-8M, with each video annotated with highlight labels based on aggregated viewing behavior data from over 50,000 viewers. This annotation method addresses the subjectivity and cost issues of traditional manual annotation, providing high-quality highlight labels that are well-suited for video highlight detection and summarization tasks.

4.2. Evaluation Metric

To evaluate the performance of our method, we employ the F-score as a metric for measuring the agreement between generated summaries and ground-truth summaries. Let X denote the generated summary for a video and Y denote the ground-truth summary. For each video, the F-score is calculated by

$$\begin{aligned} F =& \frac{2 \cdot P \cdot R}{P + R}, \\ P =& \frac{overlapped\ duration\ of\ X\ and\ Y}{duration\ of\ X}, \\ R =& \frac{overlapped\ duration\ of\ X\ and\ Y}{duration\ of\ Y}, \end{aligned} \tag{13}$$

where P and R represent the precision and recall for each video, respectively.

We also employ rank-based evaluation [29] for evaluation. It computes two rank correlation coefficients, Kendall's τ and Spearman's ρ, based on the predicted frame-level scores and the scores annotated by humans.

4.3. Implementation Details

The dimension d_m of features extracted by GoogLeNet [26] is fixed at 1024, the head number H in multi-head attention is set to 8, and the dimension of the position-wise feed-forward networks d_f is set to 2048. For the vendor stage, we use the Adam optimizer [28] with a learning rate of 0.0002 and a weight decay of 0.1. The source-only model is trained in 100 epochs with a warmup strategy [29] in the first 10 epochs. For the client stage, we set the learning rate to 0.00002 and the weight decay to 0.01 to fine-tune the source model in 100 epochs, and also use a warmup strategy in the first 10 epochs. For all settings of all datasets, we randomly divide the dataset into 5 splits and run our method five times for each setting to report the average performance of these five runs. All experiments are conducted on a single NVIDIA RTX 3090 GPU with 24 GB of graphics memory.

4.4. Quantitative Evaluation

4.4.1. Quantitative Comparison Results

We compare our method with several state-of-the-art methods under different settings, including cross-domain video summarization methods (vsLSTM [4], dppLSTM [4], DSN [10] and FPVS [9]), and standard video summarization methods (VASNet [30], SUM-FCN [1], DR-DSN [31], A-AVS [32], M-AVS [32], DSNet [2], MSVA [35], RSGN [36], 3DST-UNets$_{sup}$ [37] and RR-STG [38]).

Table 2 and 3 show the comparison results (F-score) with state-of-the-art video summarization methods on the TVSum, SumMe, and FPVSum datasets, respectively. We can have observations as follows. First, the extensive experiments on three different datasets have demonstrated the effectiveness of our method in different scenarios. Second, our method yields more favorable results compared with the cross-domain methods, which clearly demonstrates the effectiveness of the proposed discrepancy reduction loss in encouraging well alignment between the target feature distribution and the source feature distribution, as well as the effectiveness of the confidence weighting strategy in making full use of the target data. Finally, our method consistently achieves better results than the standard video summarization

Table 2: Performance comparison in terms of F-score against state-of-the-art video summarization methods on the TVSum and SumMe datasets under canonical, augmented, and transfer settings. "Adapt." indicates whether the method adapts the source model to the target domain using target unlabeled data.

Method	Venue	Adapt.	TVSum			SumMe		
			Can	Aug	Tran	Can	Aug	Tran
VASNet [30]	ACCV'18	✘	61.4	62.4	–	49.7	51.1	–
SUM-FCN [1]	ECCV'18	✘	56.8	59.2	58.2	47.5	51.1	44.1
DR-DSN [31]	AAAI'18	✘	58.1	59.8	58.9	42.1	43.9	42.6
A-AVS [32]	TCSVT'20	✘	59.4	60.8	–	43.9	44.6	–
M-AVS [32]	TCSVT'20	✘	61.0	61.8	–	44.4	46.1	–
DSNet$_{ab}$ [2]	TIP'20	✘	62.1	63.9	59.4	50.2	50.7	46.5
DSNet$_{af}$ [2]	TIP'20	✘	61.9	62.2	58.0	51.2	53.3	47.6
mvsDGCN [33]	PR'20	✘	65.0	–	–	–	–	–
DASP [34]	IJON'20	✘	63.6	64.5	–	45.5	47.0	–
MSVA [35]	ICME'21	✘	61.5	–	–	53.4	–	–
RSGN [36]	TPAMI'21	✘	60.1	61.1	60.0	45.0	45.7	44.0
3DST$_{sup}$ [37]	TIP'22	✘	58.3	58.9	56.1	47.4	49.9	47.9
RR-STG [38]	TIP'22	✘	63.0	63.6	59.7	53.4	54.8	45.4
SSPVS [39]	CoRR'22	✘	60.3	61.8	57.8	48.7	50.4	45.8
SA-CFT [40]	TIP'23	✘	62.7	60.3	58.0	56.0	54.8	44.0
VSS-Net [41]	TCSVT'24	✘	61.0	61.4	58.5	51.5	52.8	**48.4**
AMFM [42]	ESWA'24	✘	61.0	60.8	58.6	51.8	52.8	46.4
PRLVS [43]	IS'24	✘	63.0	59.2	57.0	46.3	49.7	47.6
vsLSTM [4]	ECCV'16	✔	–	–	56.9	–	–	40.7
dppLSTM [4]	ECCV'16	✔	–	–	58.7	–	–	41.8
Ours	–	✔	**65.8**	**65.9**	**60.1**	**57.4**	**57.8**	48.3

Table 3: Performance comparison in terms of F-score against state-of-the-art video summarization methods on the FPVSum datasets. "Adapt." indicates whether the method adapts the source model to the target domain using target unlabeled data.

Method	Venue	Adapt.	F-score
Random [9]	ECCV'18	✘	16.3
Uniform [9]	ECCV'18	✘	15.1
C3D [44]	ICCV'15	✘	26.9
TDCNN [45]	CVPR'16	✘	28.6
DSNet$_{ab}$ [2]	TIP'20	✘	42.1
DSNet$_{af}$ [2]	TIP'20	✘	44.6
SSPVS [39]	CoRR'22	✘	47.4
DSN [10]	NIPS'16	✔	22.7
FPVS [9]	ECCV'18	✔	35.3
Ours	–	✔	**50.3**

Table 4: Performance comparison (Kendall similarity and Spearman similarity) with state-of-the-art video summarization methods on the TVSum dataset under the canonical setting.

Method	Venue	Kendall's τ	Spearman's ρ
Random	–	0.000	0.000
Human	–	0.177	0.204
vsLSTM [4]	ECCV'16	0.042	0.055
GLRPE [46]	ECCV'20	0.070	0.091
RSGN [36]	TPAMI'21	0.083	0.090
SumGraph [47]	ECCV'20	0.094	0.138
PGL-SUM [48]	ISM'21	0.157	0.206
Clip-it [49]	NIPS'21	0.108	0.147
MSVA [35]	ICME'21	0.190	0.210
SSPVS [39]	CoRR'22	0.177	0.233
SSPVS + text [39]	CoRR'22	0.181	0.238
MFST [50]	arXiv'22	**0.222**	0.224
AAAM [51]	CVPR'23	0.193	0.254
MAAM [51]	CVPR'23	0.207	0.271
CSTA [52]	CVPR'24	0.194	0.255
Ours	–	0.212	**0.276**

Table 5: Performance comparison (Kendall similarity and Spearman similarity) with state-of-the-art video summarization methods on the SumMe dataset under the canonical setting.

Method	Venue	Kendall's τ	Spearman's ρ
Random	–	0.000	0.000
Human	–	0.177	0.204
RSGN [36]	TPAMI'21	0.083	0.085
SSPVS [39]	CoRR'22	0.178	0.240
SSPVS + text [39]	CoRR'22	0.192	0.257
MSVA [35]	ICME'21	0.200	0.230
MFST [50]	arXiv'22	0.229	0.229
AAAM [51]	CVPR'23	0.223	0.273
MAAM [51]	CVPR'23	0.227	0.278
CSTA [52]	CVPR'24	0.246	0.274
Ours	–	**0.256**	**0.286**

methods in all settings, which further verifies the effectiveness of our domain adaptation framework.

Table 4 and Table 5 further compare our method under the rank-based evaluation with other state-of-the-art methods on the TVSum and SumMe datasets respectively. The random performance is obtained by generating 100 uniformly distributed random value sequences in [0,1] for each original video and averaging the obtained correlation coefficients. The human performance is generated using the leave-one-out approach, that is, for multiple human annotations, one is selected as the prediction each time, the rest are used as labels, and the results are finally averaged. Our method generally outperforms other methods on both datasets and achieves the best results on the SumMe dataset. The excellent performance under two evaluation systems (F-score and rank-based evaluation) validates the effectiveness of our method.

Table 6: Comparison of different methods when training on the large-scale dataset Mr.HiSum and testing on TVSum and SumMe datasets, together with the test split of Mr.HiSum. * denotes our reproduced results.

Method	**Source**	**Mr.HiSum**			**TVSum**			**SumMe**		
		F-score	τ	ρ	F-score	τ	ρ	F-score	τ	ρ
SimpleMLP	Mr.HiSum	54.8	0.470	0.470	46.0	0.032	0.047	40.7	0.089	0.099
VASNet* [30]	Mr.HiSum	55.2	0.474	0.474	46.6	0.114	0.166	42.4	**0.109**	**0.120**
PGL_SUM* [48]	Mr.HiSum	55.3	0.476	0.476	47.2	0.129	0.190	42.5	0.087	0.097
SSPVS* [39]	Mr.HiSum	54.6	0.467	0.467	46.4	0.065	0.093	42.4	0.083	0.092
CSTA* [52]	Mr.HiSum	55.3	0.475	0.475	46.0	0.094	0.139	41.0	0.075	0.083
Ours	Mr.HiSum	**55.5**	**0.478**	**0.478**	**48.2**	**0.139**	**0.203**	**44.5**	0.101	0.112

To further evaluate the generalization ability of our method, we conduct domain generalization experiments using the large-scale Mr.HiSum [14] dataset. Specifically, we train our model on the Mr.HiSum training set and evaluate it on the test set of Mr.HiSum, the TVSum dataset, and the SumMe dataset. As shown in Table 6, we compare our method with the baseline (SimpleMLP), VASNet [30], and PGL_SUM [48], as well as two state-of-the-art methods (SSPVS [39] and CSTA [52]). Unlike the original Mr.HiSum paper, which adopts metrics (e.g., MAP@15, MAP@50) for highlight detection, we follow recent works and report F-score, Kendall's τ, and Spearman's ρ, which better reflect the ranking consistency of predicted and ground truth frame-level importance. Our method achieves the best performance in terms of F-score on all three datasets, showing its effectiveness in selecting representative keyframes across domains. On TVSum and Mr.HiSum, it also achieves

the highest values across all three metrics, indicating strong robustness and ranking quality.

4.4.2. Ablation Study

To perform an in-depth analysis of each individual component of our method, we conduct extensive ablation studies on TVSum, SumMe and FPVSum datasets. We replace our generalized Transformer with the standard Transformer ("w/o TF") to evaluate its impact on the complete model. We remove the discrepancy reduction loss ("w/o DRL") and the confidence weighting strategy ("w/o CWS"), respectively, to evaluate their individual effect on the domain adaptation. Table 7 shows the ablation study results. It is interesting to observe that the proposed generalized Transformer has better overall performance than the standard Transformer. Moreover, it is obvious that both the discrepancy loss and the confidence weighting strategy are beneficial for improving the performance under all settings on all datasets.

Table 7: Ablation study results (F-score) on the TVSum, SumMe and FPVSum datasets.

Method	TVSum			SumMe			FVPSum
	Can	Aug	Tran	Can	Aug	Tran	
w/o TF	63.2	64.9	58.7	48.4	55.3	45.5	47.2
w/o DRL	65.3	65.0	59.6	52.7	55.1	45.2	46.5
w/o CWS	65.7	65.7	59.9	56.7	56.5	47.4	49.4
Ours	**65.8**	**65.9**	**60.1**	**57.4**	**57.8**	**48.3**	**50.3**

Table 8: Comparison results (Maximum Mean Discrepancy) between standard Transformer and our generalized Transformer on the TVSum, SumMe and FPVSum datasets at the vendor stage.

Method	TVSum			SumMe			FVPSum
	Can	Aug	Tran	Can	Aug	Tran	
Standard	0.12	0.09	0.08	0.28	0.33	0.28	0.41
Generalized	**0.08**	**0.08**	**0.07**	**0.22**	**0.29**	**0.18**	**0.39**

In order to further analyze the effectiveness of our method, we conducted experiments at two different stages. As shown in Table 8, we use Maximum Mean Discrepancy between the source and target domains to evaluate the generalization ability of the model and compare the performance of

the standard Transformer and our generalized Transformer at the vendor stage. Maximum Mean Discrepancy (MMD) is a widely used metric in domain adaptation [53] and domain generalization [54], which quantitatively measures the discrepancy between two distributions. Specifically, MMD calculates the difference between the empirical means of the source and target feature distributions in a reproducing kernel Hilbert space (RKHS). It is interesting to note that our generalized Transformer has a smaller Maximum Mean Discrepancy, which represents a greater generalization ability than the standard Transformer.

At the client stage, as shown in Table 9, we add our discrepancy reduction loss and confidence weighting strategy to two existing methods (VASNet [30], DR-DSN [31], DSNet [2] and PGL-SUM [48]), and compare them with their original versions. We can observe that the methods achieve better results than their original versions, indicating that the strategy at the client stage can work independently without the generalized Transformer model, which has strong practicality.

Table 9: Comparison results (F-score) between the original methods and those that additionally applied our domain adaptation strategy on the TVSum and SumMe datasets at the client stage.

Method	TVSum			SumMe		
	Can	Aug	Tran	Can	Aug	Tran
VASNet [30]	61.4	62.4	–	49.7	51.1	–
VASNet + Ours	**61.9**	**63.1**	–	**51.4**	**52.0**	–
DR-DSN [31]	58.1	59.8	58.9	42.1	43.9	42.6
DR-DSN + Ours	**59.4**	**60.8**	**59.4**	**43.5**	**45.1**	**43.6**
DSNet$_{af}$ [2]	61.9	62.2	58.0	51.2	53.3	47.6
DSNet$_{af}$ + Ours	**62.3**	**63.0**	**59.0**	**51.9**	**54.2**	**48.2**
PGL-SUM [48]	61.0	–	–	55.6	–	–
PGL-SUM + Ours	**61.5**	–	–	**57.3**	–	–

4.5. Qualitative Evaluation

4.5.1. Qualitative Comparison Results

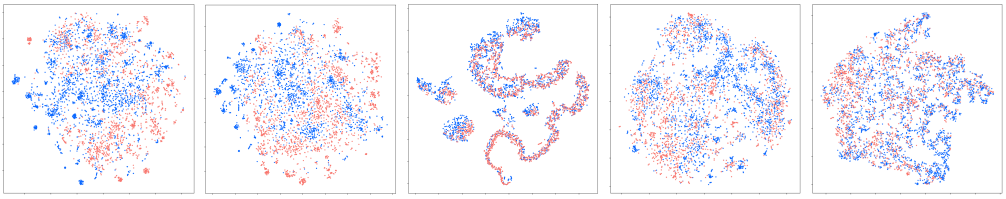
Figure 3 provides example summaries of several videos generated by different methods (the ground-truth, our method, "Generalized", CSTA [52] and SSPVS [39]), including the 2-nd, 14-th and 43-rd videos from the TVSum dataset under the transfer setting, where "Generalized" represents directly applying the generalized Transformer to the target domain without

adaptation. We observe that the summaries generated by our method have a high overlap with the ground-truth summaries, and our method outperforms the generalized Transformer and other methods, which demonstrates the effectiveness of our method in reducing domain shift.

Figure 3: Qualitative results of different video summarization methods. The colored line segments denote the selected video segments for the corresponding method, and the frames below are sampled from the ground-truth summaries.

4.5.2. Visualization Results

To intuitively explain the effect of our method, we compare the visualization of frame feature distribution between our method and other methods, including standard Transformer, SSPVS [39], CSTA [52], and the variant of our method (generalized Transformer). Specifically, we plot the T-SNE visualization of features of source and target domains under the transfer setting on the TVSum dataset, where the feature distribution of the target domain is different from the source domain. It is worth noting that our goal is to assess



(a) Standard (b) SSPVS (c) CSTA (d) Generalized (e) Ours

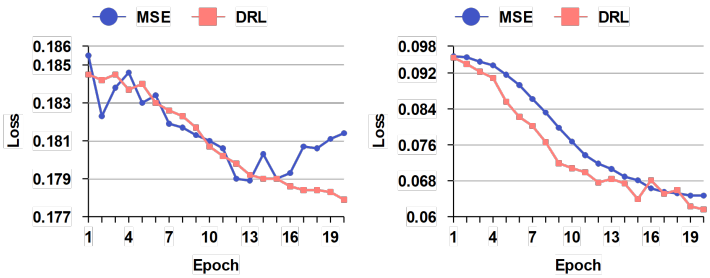
Figure 4: T-SNE visualizations of frame feature distributions generated by the standard Transformer, SSPVS [39], CSTA [52], the generalized Transformer, and our method, in which different colors are used to distinguish the frame features belonging to different domains.

how well features from different domains are aligned, rather than focusing solely on the compactness of clusters within a single domain. As shown in Figure 4, the generalized Transformer learns more consistent features between the source and target domains, and our method further improves the extraction of domain-invariant representations, outperforming the standard Transformer, SSPVS, CSTA, and the generalized Transformer.

Figure 5 presents the change in both discrepancy reduction loss (red) and MSE loss (blue) during the client stage. The discrepancy reduction loss is used to adapt the source model to the target domain using unlabeled target data. The MSE loss, computed between the predicted and ground-truth importance scores, serves as a performance metric on the target domain. Figure 5 (a) and (b) correspond to the transfer settings on SumMe and TVSum, respectively. Notably, in both plots, the MSE loss decreases in sync with the discrepancy reduction loss, indicating that minimizing the discrepancy reduction loss effectively helps align the target feature distribution with the source feature distribution, leading to more accurate predictions. Moreover, compared to Figure 5 (b), the MSE loss curve in Figure 5 (a) exhibits minor fluctuations, likely due to the smaller number of annotated training samples in the SumMe dataset.

4.6. Complexity Analysis

We further compare the time complexity and memory usage of our method with the state-of-the-art Transformer-based methods and non-Transformer-based methods. We measure several metrics, including training latency, total training time, training memory, inference latency and inference memory. As shown in Table 10, compared to methods that do not use an attention mech-



(a) SumMe dataset (b) TVSum dataset

Figure 5: Loss change curves on the transfer setting of SumMe and TVSum, where the MSE loss for evaluation is shown in blue and the discrepancy reduction loss (DRL) is shown in red.

anism, such as DR-DSN, which relies on LSTM [19], our method achieves better inference efficiency and shorter training cost. In addition, although we adopt a two-stage training strategy, the total training time is still shorter than DSNet$_{ab}$, DSNet$_{af}$ and CSTA. These results demonstrate that the proposed method maintains a favorable trade-off between accuracy and computational cost.

Table 10: Model complexity comparison between our method and other methods.

Method	**Params**	**Training**			**Inference**	
		Latency (ms/video)	Memory (GB)	Total Time (s)	Latency (ms/video)	Memory (GB)
DR-DSN [31]	2.6 M	36.4	0.66	248.9	41.7	0.40
DSNet$_{ab}$ [2]	4.3 M	13.1	1.04	447.0	1.8	0.56
DSNet$_{af}$ [2]	4.3 M	13.7	0.88	468.5	1.3	0.55
CSTA [52]	8.4 M	636.2	2.59	7252.7	19.3	0.93
SSPVS [39]	37.8 M	71.1	1.92	81.1	4.1	0.99
Ours (Vendor)	7.5 M	7.9	1.31	90.1	–	–
Ours (Client)	7.5 M	15.2	1.88	173.3	–	–
Ours	7.5 M	23.1	1.88	263.4	2.8	0.61

5. Limitation and Future Work

Despite the promising performance of our method, it is built upon the assumption that the similarity distribution of visual features between video frames aligns with the similarity distribution of their corresponding importance scores. That is, if two frames are visually similar in feature space, their

importance values are likely to be similar. While this assumption holds true in many cases, it may not always be valid, especially when other modalities such as audio or text contribute significantly to the importance of frames.

In future work, we are going to explore multi-modal domain adaptation techniques that incorporate audio and textual cues to achieve better summarization. We believe that integrating such complementary information can enhance the robustness and generalization ability of our summarization framework, especially in scenarios where visual features alone are not sufficient to capture the saliency of the content.

6. Conclusion

In this work, we address the challenge of cross-domain video summarization, a task that suffers from the domain shift between the source and target video domains. This domain shift often leads to poor performance when applying a model trained on one domain to videos from a different domain. To alleviate this problem, we propose a novel domain adaptation framework consisting of two stages. At the vendor stage, we introduce a generalized Transformer model, in which a regularization strategy is applied to the attention mechanism to enhance the model's generalization ability across diverse video domains. To further improve domain adaptation, at the client stage, we design a discrepancy reduction loss and a confidence weighting strategy, which together ensure that the model adapts well to the target domain while maintaining robust and transferable representations. Experimental results demonstrate that our method outperforms existing methods in both generalization and adaptation across multiple domains.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No 62072041.

References

[1] M. Rochan, L. Ye, Y. Wang, Video summarization using fully convolutional sequence networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 347–363.

[2] W. Zhu, J. Lu, J. Li, J. Zhou, Dsnet: A flexible detect-to-summarize network for video summarization, IEEE Transactions on Image Processing 30 (2020) 948–962.

[3] W. Zhu, J. Lu, Y. Han, J. Zhou, Learning multiscale hierarchical attention for video summarization, Pattern Recognition 122 (2022) 108312.

[4] K. Zhang, W.-L. Chao, F. Sha, K. Grauman, Video summarization with long short-term memory, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2016, pp. 766–782.

[5] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2010, pp. 213–226.

[6] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2066–2073.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: International Conference on Machine Learning (ICML), PMLR, 2014, pp. 647–655.

[8] B. Sun, J. Feng, K. Saenko, Return of frustratingly easy domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 30, 2016.

[9] H.-I. Ho, W.-C. Chiu, Y.-C. F. Wang, Summarizing first-person videos from third persons' points of view, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 70–85.

[10] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, Domain separation networks, Advances in neural information processing systems 29 (2016).

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[12] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, Tvsum: Summarizing web videos using titles, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5179–5187.

[13] B. Gong, W.-L. Chao, K. Grauman, F. Sha, Diverse sequential subset selection for supervised video summarization, Advances in neural information processing systems 27 (2014) 2069–2077.

[14] J. Sul, J. Han, J. Lee, Mr. hisum: A large-scale dataset for video highlight detection and summarization, Advances in Neural Information Processing Systems 36 (2023) 40542–40555.

[15] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, A. de Albuquerque Araújo, Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognition Letters 32 (1) (2011) 56–68.

[16] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, B. E. Ionescu, Video summarization from spatio-temporal features, in: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, 2008, pp. 144–148.

[17] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, H. Zha, Unsupervised deep learning for optical flow estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 31, 2017.

[18] B. Zhao, X. Li, X. Lu, Hierarchical recurrent neural network for video summarization, in: Proceedings of the 25th ACM international conference on Multimedia (MM), 2017, pp. 863–871.

[19] B. Zhao, X. Li, X. Lu, Tth-rnn: Tensor-train hierarchical recurrent neural network for video summarization, IEEE Transactions on Industrial Electronics 68 (4) (2020) 3629–3637.

[20] S.-H. Zhong, J. Lin, J. Lu, A. Fares, T. Ren, Deep semantic and attentive network for unsupervised video summarization, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 18 (2) (2022) 1–21.

[21] I. Pattnaik, P. Narwal, Implicit embedding based multi modal attention network for cricket video summarization, Engineering Applications of Artificial Intelligence (EAAI) 148 (2025) 110428.

[22] P. Narwal, N. Duhan, K. Kumar Bhatia, A comprehensive survey and mathematical insights towards video summarization, Journal of Visual Communication and Image Representation (JVCIR) 89 (2022) 103670.

[23] P. Narwal, N. Duhan, K. K. Bhatia, A novel multi-modal neural network approach for dynamic and generic sports video summarization, Engineering Applications of Artificial Intelligence (EAAI) 126 (2023) 106964.

[24] P. Narwal, N. Duhan, K. K. Bhatia, Domain knowledge based multi-cnn approach for dynamic and personalized video summarization, in: International Conference on Communication and Computational Technologies (ICCCT), Springer, 2024, pp. 81–94.

[25] H. Jiang, Y. Mu, Joint video summarization and moment localization by cross-task sample transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16388–16398.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[27] Open video project, http://www.open-video.org/.

[28] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[29] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2017.

[30] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, P. Remagnino, Summarizing videos with attention, in: Computer Vision – ACCV 2018 Workshops, Springer, 2019, pp. 39–54.

[31] K. Zhou, Y. Qiao, T. Xiang, Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 32, 2018.

[32] Z. Ji, K. Xiong, Y. Pang, X. Li, Video summarization with attention-based encoder–decoder networks, IEEE Transactions on Circuits and Systems for Video Technology 30 (6) (2019) 1709–1717.

[33] J. Wu, S. hua Zhong, Y. Liu, Dynamic graph convolutional network for multi-video summarization, Pattern Recognition (PR) 107 (2020) 107382.

[34] Z. Ji, F. Jiao, Y. Pang, L. Shao, Deep attentive and semantic preserving video summarization, Neurocomputing (IJON) 405 (2020) 200–207.

[35] J. A. Ghauri, S. Hakimov, R. Ewerth, Supervised video summarization via multiple feature sets with parallel attention, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6s.

[36] B. Zhao, H. Li, X. Lu, X. Li, Reconstructive sequence-graph network for video summarization, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (5) (2021) 2793–2801.

[37] T. Liu, Q. Meng, J.-J. Huang, A. Vlontzos, D. Rueckert, B. Kainz, Video summarization through reinforcement learning with a 3d spatio-temporal u-net, IEEE Transactions on Image Processing 31 (2022) 1573–1586.

[38] W. Zhu, Y. Han, J. Lu, J. Zhou, Relational reasoning over spatial-temporal graphs for video summarization, IEEE Transactions on Image Processing 31 (2022) 3017–3031.

[39] H. Li, Q. Ke, M. Gong, R. Zhang, Video summarization based on video-text modelling, CoRR, abs/2201.02494 2 (7) (2022) 8.

[40] R. Zhong, R. Wang, W. Yao, M. Hu, S. Dong, A. Munteanu, Semantic representation and attention alignment for graph information bottleneck in video summarization, IEEE Transactions on Image Processing 32 (2023) 4170–4184.

[41] Y. Zhang, Y. Liu, W. Kang, R. Tao, Vss-net: Visual semantic self-mining network for video summarization, IEEE Transactions on Circuits and Systems for Video Technology 34 (2024) 2775–2788.

[42] Y. Zhang, Y. Liu, C. Wu, Attention-guided multi-granularity fusion model for video summarization, Expert Systems with Applications (ESWA) 249 (2024) 123568.

[43] G. Wang, X. Wu, J. Yan, Progressive reinforcement learning for video summarization, Information Sciences (IS) 655 (2024) 119888.

[44] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2015, pp. 4489–4497.

[45] T. Yao, T. Mei, Y. Rui, Highlight detection with pairwise deep ranking for first-person video summarization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 982–990.

[46] Y. Jung, D. Cho, S. Woo, I. S. Kweon, Global-and-local relative position embedding for unsupervised video summarization, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2020, pp. 167–183.

[47] J. Park, J. Lee, I.-J. Kim, K. Sohn, Sumgraph: Video summarization via recursive graph modeling, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2020, pp. 647–663.

[48] E. Apostolidis, G. Balaouras, V. Mezaris, I. Patras, Combining global and local attention with positional encoding for video summarization, in: 2021 IEEE international symposium on multimedia (ISM), IEEE, 2021, pp. 226–234.

[49] M. Narasimhan, A. Rohrbach, T. Darrell, Clip-it! language-guided video summarization, Advances in neural information processing systems 34 (2021) 13988–14000.

[50] J. Park, K. Kwoun, C. Lee, H. Lim, Multimodal frame-scoring transformer for video summarization, arXiv preprint arXiv:2207.01814 (2022).

[51] H. Terbouche, M. Morel, M. Rodriguez, A. Othmani, Multi-annotation attention model for video summarization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3143–3152.

[52] J. Son, J. Park, K. Kim, Csta: Cnn-based spatiotemporal attention for video summarization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18847–18856.

[53] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, W. Zuo, Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2272–2281.

[54] T. Nguyen, B. Lyu, P. Ishwar, M. Scheutz, S. Aeron, Joint covariate-alignment and concept-alignment: a framework for domain generalization, in: 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2022, pp. 1–6.