Video Sampled Frame Category Aggregation and Consistent Representation for Cross-Modal Retrieval

Ming Jin, Huaxiang Zhang, Lei Zhu, *Senior Member, IEEE*, Jiande Sun, and Li Liu

***Abstract*— Many current video and text cross-modal retrieval research works focus on narrowing the semantic gap between video and text, but ignore the semantic difference between different sampled frames in the same video and the correlation of feature distribution of objects contained in different sampled frames in the same video, as a result, the features of the sampled frames in the final learned video cannot well represent the semantic features of the whole video. To overcome the shortcomings of existing studies, we first use a pre-trained video frame classification-aggregation network to make the object categories contained in different sampled frames in the same video be more close to the important object categories contained in the whole video, so as to promote the feature distribution of different sampled frames in the same video to be consistent, and increase the relevance of object features in different frames. Then we propose a video internal frame aggregation loss module to solve the problem of inconsistent feature distribution between different frame features encoded by video encoder in the same video and the aggregation feature of the sampled frame, thus enhancing the ability of video sampled frame aggregation feature representation. Experiments conducted on three common datasets MSVD, MSR-VTT and DiDeMo demonstrate the validity of the proposed approach.**

***Index Terms*— Video and text cross-modal retrieval, pre-trained video frame classification-aggregation network, video internal frame aggregation loss module.**

I. Introduction

With the development of Internet technology, people's access to information has undergone earth-shaking changes, especially the access to video information on the Internet. This promotes the transformation of retrieval technology from the original single modal information retrieval (text retrieval text, image retrieval image) to multi-modal retrieval (video and text cross-modal retrieval, image and text cross-modal retrieval). With the increase of people's demand, video and text cross-modal retrieval technology also ushers in a boom of research in recent years. A straight view of video and text cross-modal retrieval is shown in Fig. 1. There are two main reasons for the research upsurge of video and text cross-modal retrieval: one is the demand of video and text retrieval technology, and other is that, image and text cross-modal retrieval technology [1], [2] has been mature, some researchers do further research by transferring technology to video and text cross-modal retrieval tasks. The research of video and text cross-modal retrieval can be divided into two major directions, one is based on the non-pre-trained model [3], [4], [5], [6], [7] and the other is based on the pre-trained transfer model [8], [9], [10]. In the early stages of video and text cross-modal retrieval, researchers used a non-pre-trained model [11]. The video and text cross-modal retrieval network architecture of the non-pre-trained model is basically similar to that of the image and text cross-modal retrieval network. This network architecture achieves good performance in image and text cross-modal retrieval, but it does not perform well in video and text cross-modal retrieval. Each video consists of multiple static frames with sequential sequence, and each static frame can be regarded as an image, so the amount of information in a video is much larger than that in an image, the conventional network model cannot adequately capture the information contained in the video. However, the network architecture of image feature processing is helpful to that of video feature processing. In image and text cross-modal retrieval, the advantage of

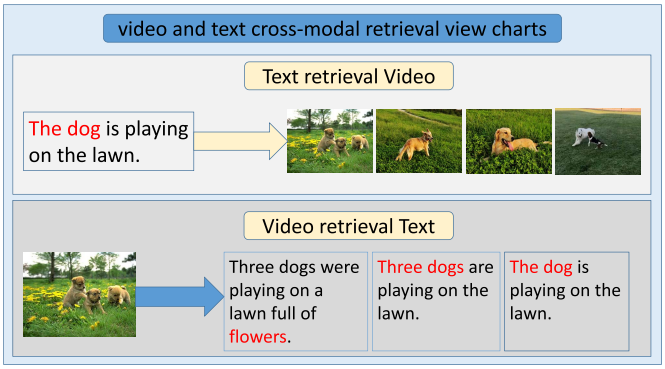


Fig. 1. The video and text cross-modal retrieval view charts.

Manuscript received 23 April 2022; revised 2 August 2022; accepted 4 September 2022. Date of publication 5 September 2022; date of current version 6 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62176144, Grant U1836216, and Grant 62076153; in part by the Major Fundamental Research Project of Shandong, China, under Grant ZR2019ZD03; and in part by the Taishan Scholar Project of Shandong, China, under Grant ts20190924. This article was recommended by Associate Editor F. Díaz-de-María. *(Corresponding authors: Huaxiang Zhang; Li Liu.)*

Ming Jin, Lei Zhu, Jiande Sun, and Li Liu are with the School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China.

Huaxiang Zhang is with the School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China, and also with the School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan 250357, China (e-mail: huaxzhang@hotmail.com; liuli_790209@163.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3204623.

Digital Object Identifier 10.1109/TCSVT.2022.3204623

image processing network is that it can mine the spatial information of image features. Therefore, using image feature processing network to extract the static frame features of each video can fully establish the spatial information of frame features. The video and text cross-modal retrieval models have taken the advantages of image processing network design in image and text cross-modal retrieval models. By getting some inspiration from the image feature processing network, some researchers have applied image feature processing network to video frame feature processing [9]. They first use a large number of images to train the image feature processing network, and obtain the stable state network parameters, then, they transfer the image feature processing network to the video frame feature processing network. Considering that the frame features of video and image features have certain differences, therefore, the transported model should be fine-tuned during the training process.

In order to establish the temporal information between the sampled frames in the video, a non-pre-trained 3D convolution is also used to extract video features, which can not only solve the spatial information contained in the video features, but also solve the temporal information in the video features. Tran *et al.* [12] used 3D convolution to extract video features in early video feature extraction. The spatial and temporal information of features can be mined by using 3D convolution. Although the use of 3D convolution has improved the quality of video feature extraction, the requirements on hardware equipment have been improved, therefore, many researchers still apply 2D convolution to extract video features. Extracting video features using 2D convolution is to extract the features of video sampled frames, so the video feature processing is transformed into the frame feature processing of video samples. Some researchers gradually train image feature and text feature processing models on large-scale image datasets, and then transfer the pre-trained models to the video frame feature and text feature processing network by initializing them with the parameters of the pre-trained models. Contrastive Language-Image Pre-training (CLIP) [8] is a representative and effective model for image and text feature processing. Therefore, in our video and text cross-modal retrieval model, the image encoder and text encoder of CLIP are applied to the backbone network of video and text cross-modal retrieval. Although the pre-trained model is gradually used in cross-modal research, the existing video and text cross-modal retrieval model based on pre-trained model lacks consideration of the relationship between video frames. As we know, the final feature representation of a video is determined by the features of multiple sampled frames in the video, which highlights the importance of the relationship between frames in the video. Although previous studies on spatial and temporal information of video feature have been carried out, the research on the relationship between video sampled frames is still lacking. To make up for this lack, we propose a video frame classification-aggregation network, which uses the clustering idea to fully mine the key object features in video modality. The feature categories extracted from each frame in the video are aggregated to the key groups of the whole video. The simulation diagram of the video frame classification-aggregation network is shown in Fig. 2. In Fig. 2, the training process of neural network is represented from left to right. With the training of neural network, important object features of different sampled frames in the same video become more prominent, and unimportant object features in each frame are gradually abandoned or weakened. Secondly, we propose a video internal frame aggregation loss module to improve the semantic consistency between the video aggregation feature and the sampled frame feature in the video.

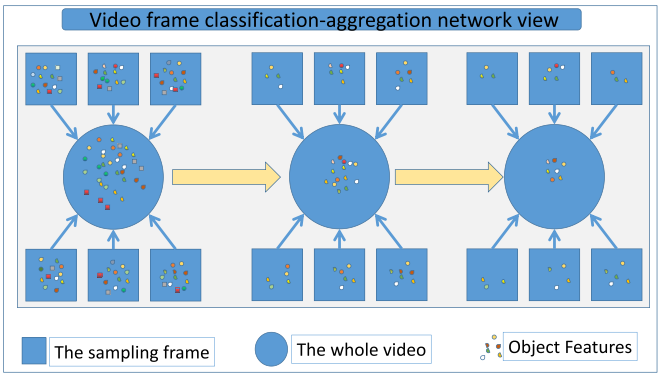


Fig. 2. The simulation diagram of the video frame classification-aggregation network.

Our contributions are as follows:

(1) The application of video frame classification-aggregation network makes important object features contained in different sampled frames in the video more prominent, while unimportant object features are gradually weakened or abandoned. Thus, the problem of alignment difficulty between video modality and text modality in training network is solved.

(2) The design of video internal frame aggregation loss module can reduce the difference of feature distribution between the finally learned video aggregation features and the features of different video sampled frames. Through continuous learning of the network, the features of each frame can better represent the features of the important objects contained in the video.

(3) Pre-trained model is applied to cross-modal retrieval and can well learn the semantic space of video modality and text modality. It greatly reduces the training time of neural network in semantic space learning and improves the retrieval performance of the model. We verify the effectiveness of the model on three datasets MSVD, MSR-VTT and DiDeMo.

II. Related Work

A. Image and Text Cross-Modal Retrieval

The development of image and text cross-modal retrieval technology has greatly promoted the development of video and text cross-modal retrieval technology. The difference between them lies in the video and image feature processing, but the design of image feature processing network has a great guiding effect on the design of video feature processing network, therefore, it is necessary to have a comprehensive understanding of image and text cross-modal retrieval technology

before studying video and text cross-modal retrieval technology. In cross-modal retrieval, feature extraction develops from coarse-grained features to fine-grained features. Fine-grained features can better represent the potential semantic features of things. Among them, He *et al.* [13] proposed a fine-grained cross-modal retrieval benchmark and method. The benchmark and method contain three constraints (classification constraint, central constraint, ranking constraint), which can enhance the learning of discriminating features and make the features more compact. Yu *et al.* [14] extended single-modal retrieval to multi-modal retrieval by reordering method, realizing a bidirectional coarse-to-fine cross-modal retrieval structure, thus improving cross-modal retrieval performance. In order to pay attention to the interaction between image regions and sentence words, and select the most significant information, Wang *et al.* [15] utilized global and local cross-modal interactive alignment methods to fully align the global and local features of both image and text, and used adaptive gating methods to process irrelevant information. Both coarse-grained feature and fine-grained feature extraction take a lot of time to label images, moreover, simple label information cannot adequately represent the semantics contained in the image. Therefore, unsupervised cross-modal retrieval has been extensively explored. Peng *et al.* [16] made a further study on unsupervised cross-modal retrieval methods. The model uses the scene graph adaptive method to transfer the knowledge learned from the source domain to the target domain, so as to improve the adaptive ability of the target domain. Processing a large number of high-dimensional image features not only takes up a lot of storage space, but also takes a lot of time to do operations. To compensate for this deficiency, Zhu *et al.* [1] designed an image retrieval model using deep hashing. This model uses dual-level Semantic Transfer to learn hash representation and mines latent semantic information in tags. Applying hashing learning to image and text feature representation greatly improves the retrieval efficiency. In supervised cross-modal retrieval, Qian *et al.* [17] constructed a dual adversarial graph neural network to learn common representations of modal invariance and discrimination. In this model, a multi-hop neural network with a layer aggregation mechanism is used to propagate information, so as to obtain the correlation of labels and the classifier of learning mutual dependence. Dong *et al.* [18] emphasized the importance of the integrity of semantic representation in cross-modal retrieval, therefore, and used graph convolution network in their model to establish the relationship between the sample itself and the neighborhood to learn sample representation. In order to promote the alignment of abstract semantic features between different modalities, Dong *et al.* [19] also proposed a hierarchical aggregation algorithm based on graph convolution network. The algorithm generates specific modal features by integrating object features and global features of other modalities. The essence of considering the relationship between fine-grained features of different modalities is to align the features of fine-grained objects of different modalities. Lee *et al.* [20] utilized the stacked cross attention mechanism to establish a corresponding relationship between image regions and words in sentences, so as to highlight the semantic alignment between important objects in images and things described in texts, thus narrowing the semantic differences between different modalities. Excessive alignment fine-grained features of image areas and text may misunderstand the underlying semantics of the image. Therefore, the relationship between image fine-grained features becomes more and more important to the representation of image semantic features. Among them, Wu *et al.* [2] used the regional reinforcement network to explore the relationship between the image fine-grained regional features, and then aligned the text fine-grained features. Hu *et al.* [21] used a dual attention network to establish a corresponding relationship between image regions and text words, and explore the inherent abstract semantic relationship between image regional features and text features. When fine-grained features are aligned, it means that different modal features can be fully aligned. Peng *et al.* [22] proposed multi-level adaptive visual-textual alignment method, which can not only make adaptive alignment and connection between image and text from global and local perspectives, but also can align image regions and and their corresponding text words.

The design of image and text cross-modal retrieval model is a process from coarse-grained features extraction network to fine-grained features extraction network design. Both the mutual retrieval between coarse-grained features of different modalities and the mutual retrieval between fine-grained features of different modalities are the alignment of target features of different modalities.

B. Video and Text Cross-Modal Retrieval

Video and text cross-modal retrieval technology has developed from non-pre-trained model to pre-trained model. Video and text cross-modal retrieval technology based on pre-trained model has achieved a leap in retrieval accuracy, which provides new ideas for the recent research on video and text cross-modal retrieval. Previous video and text cross-modal retrieval models need to be trained on large video datasets, which is limited by the size of video datasets. Video datasets are more difficult to process than image datasets because video contains more information. It takes a lot of time to generate video corresponding text description in video dataset. The quality of text description generated directly from video is not very good. In order to alleviate the limitation of lack of training data, researchers carried out training on image dataset, and transferred the trained network model. The transferred models are trained and fine-tuned on smaller video datasets. As we know, video and text belong to different modalities, and retrieval between different modalities should overcome semantic differences between them. Miech *et al.* [3] proposed an embedded hybrid expert model, which can learn from heterogeneous data sources to deal with the missing video modalities during pre-training. The model uses image mapping data to enhance video subtitle dataset during training. Yu *et al.* [4] used a joint sequence fusion model to combine the feature representations of two paired sequence data into a 3D tensor. A hierarchical decoder is used to establish the potential semantic relationship between different levels of two modalities, and the similarity measurement is carried out.

The hierarchical measurement mechanism can effectively promote the similarity matching between different modalities and remove the irrelevant feature representation from top to bottom. The top-down similarity measurement method focuses on different targets at different measurement layers, and each layer filters out some irrelevant data in the two modalities, thus achieving accurate measurement. Zhang *et al.* [5] applied the idea of hierarchical modeling to the design of cross-modal retrieval model. In this model, global feature alignment and local feature alignment of video feature and text feature are placed at different levels. By establishing global and local alignment loss at different levels, the consistent expression of the two modal features is promoted. The early design of video and text cross-modal retrieval models mainly focus on aligning video and text features, and then design different loss functions to reduce the semantic differences between the two modalities. The semantic gap between different modal features not only exists in video (image) and text cross-modal retrieval, but also in video time location. Hu *et al.* [23] proposed an end-to-end cross-modal hash coding network. The network projects text coding and video coding into a common Hamming space, and calculates their similarity by hamming distance, so as to realize video location function. Projecting the hash codes of two different modalities into a common Hamming space further reduces the semantic differences between the two modalities. There is semantic gap between different modalities and inconsistent semantic distribution between different video frames. To solve this problem, Qi *et al.* [24] proposed a binary representation learning framework, and applied a spatial-temporal context and semantic relations of different frames to solve the semantic inconsistency in cross-modal video retrieval. Therefore, semantic gap is the key issue of cross-modal retrieval research. With the continuous optimization of transformer network, transformer network is gradually applied in video feature processing and text feature processing. Among them, Ging *et al.* [25] performed a cooperative hierarchical transformer network, in which the video feature processing branch and the text feature processing branch are symmetrical. The model uses hierarchical information between different modal features to model interactions between different granularity levels and different modalities. The temporal information between video frames has also been further explored. Dong *et al.* [26] designed a symmetric network model by using the method of hybrid semantic space learning. With the increase of network depth, the model extracts coarse-grained features and fine-grained features of video and text, and fuses them in a hybrid semantic space. Compared with other models, this model uses multi-layer GRU network to extract temporal information of video modality and text modality.

C. Pre-Trained Modal

Neural networks need to be trained through a large amount of data to learn effective feature representation. However, existing video datasets are relatively small, and some researchers pre-train video encoders on a large number of image datasets, and then train and fine-tune video encoders on video datasets. In the video and text cross-modal retrieval, pre-trained models are mainly divided into two types: one is a single pre-trained video feature extraction network, and the other is a multi-modal joint pre-trained feature extraction network. In the single pre-trained video feature extraction network, the pre-trained network extracts and stores video features. Researchers then design cross-modal network structures to process the stored video features. Among them, Liu *et al.* [6] executed a collaborative expert model to aggregate different modal features of video to represent the whole video feature. Gabeur *et al.* [27] proposed a pre-trained video feature extractor (expert) to extract different modal features contained in the video. Different from the collaborative expert model, this model uses the idea of hierarchical measurement of different video modal features to fully mine the information contained in different video modalities. In the above two methods, the video feature extraction network does not participate in the training process of the neural network. However, in the multi-modal joint pre-trained feature extraction network, the visual feature extraction network and text feature extraction network are used for joint training. Then, the visual and text feature jointly pre-trained extraction network is applied to the design of video and text cross-modal retrieval network structure. Such as, Radford *et al.* [8] proposed the Contrastive Language-image pre-training (CLIP) model, which uses natural language and matched image contrast learning to generate a pre-trained model on large-scale datasets. The pre-trained CLIP can be used to predict whether an image and a text are a pair. Portillo-quintero *et al.* [9] also applied CLIP to video retrieval to learn the public space of image and text, and then transferring it to the video dataset to learn the public space of video and text. This model is a preliminary exploration of transferring the pre-trained model of image and text cross-modal retrieval to video and text cross-modal retrieval model. We also use this idea in the model design and continue to make some progress. As we know, the selection of video frame plays a decisive role in video feature extraction. Among them, Lei *et al.* [10] utilized the method of sparse sampling to sample video frames, and the sampled frames are processed by pre-trained BERT. This model compares the features of each sampled frame in the video with the features of the text, rather than averaging the features of the sampled video frame. The model can fully explore the important target features of video frames. The main difference between the single pre-trained video feature extraction network and the multi-modal joint pre-trained feature extraction network is that the multi-modal joint pre-trained feature extraction network learns the semantic space of visual and text in advance through pre-training, and the network further narrows the semantic difference between video modality and text modality through joint training with the subsequent cross-modal retrieval network. For the single pre-trained video feature extraction network, the network does not participate in the subsequent cross-modal retrieval network training, and there is still a large semantic gap between the pre-stored video features and the extracted text features. Pre-trained models have become a trend for video feature processing, which can not only improve the retrieval accuracy but also reduce the training time on video datasets.

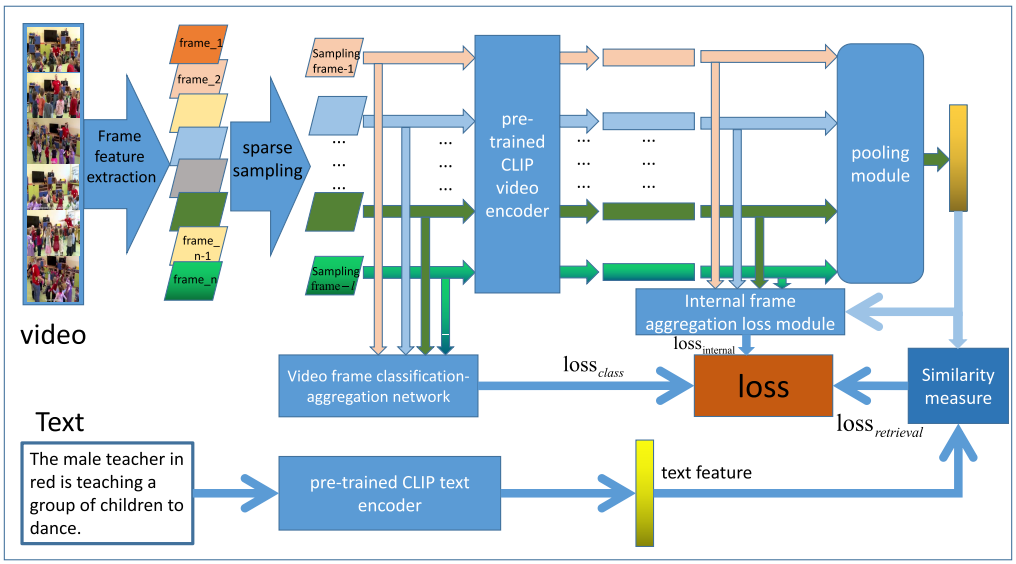


Fig. 3. The overall architecture diagrams for video and text cross-modal retrieval.

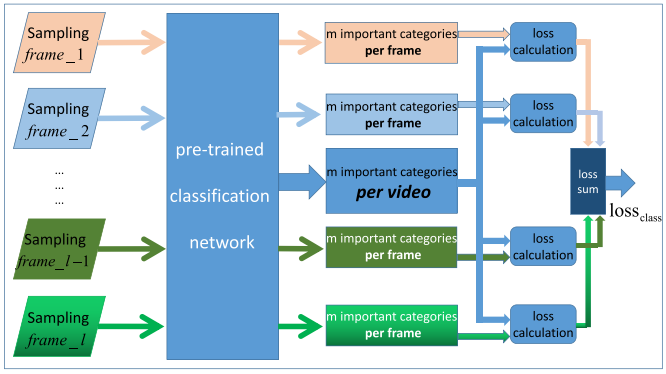


Fig. 4. The architecture diagram of the video frame classification-aggregation network.

III. Our Method

A. Video Feature Processing Network Branch

The branch of video feature processing network is mainly composed of three parts, which are video frame classification-aggregation network, video frame feature coding network and video internal frame aggregation loss module. The overall architecture of video and text cross-modal retrieval is shown in Fig. 3. The model's data loader reads the video feature $\mathbf{V} = \{\mathbf{v}_i \in \mathrm{R}^{l\times c\times n\times n}\}_{i=1}^{b}$ in batches. Where, b is the number of videos in each batch, l represents the number of sampled frames in each video, c represents the number of channels in each frame, and n represents the size of frame.

1) Video Frame Classification-Aggregation Network: The architecture diagram of the video frame classification-aggregation network is shown in Fig. 4. The video frame classification-aggregation network uses the pre-trained $EfficientNet$ network $\mathrm{E}(\mathbf{V})$ [28], and the batch video feature $\mathbf{V}$ is input into the $EfficientNet$ network $\mathrm{E}(\mathbf{V})$ to obtain the classification $\mathbf{K} \in \mathrm{R}^{b\times l\times h}$ of each frame in the video. Where h is the number of categories. We screened out the category $\mathbf{K}_{frame\_max} \in \mathrm{R}^{b\times l\times m}$ with high probability of occurrence in each frame (m is the number of categories with high probability of occurrence in each frame). In addition, category $\mathbf{K}_{video\_max} \in \mathrm{R}^{b\times m}$ with high probability of occurrence of each video is screened. By calculating the categories of each frame and the significant categories contained in the whole video, we can figure out the $\mathrm{loss}_{\mathrm{class}}$. The pseudocode for solving $\mathrm{loss}_{\mathrm{class}}$ is shown in Algorithm 1.

Algorithm 1 Class Loss Calculation

input : Batch video features: $\mathbf{V} \in \mathrm{R}^{b\times l\times c\times n\times n}$
output: $\mathrm{loss}_{\mathrm{class}}$

Pre-trained classification network: $\mathbf{K} \in R^{b\times l\times h}$;
m class index of maximum probability per frame: $\mathbf{K}_{frame\_max} \in R^{b\times l\times m} \leftarrow top\_m(\mathbf{K} \in R^{b\times l\times h})$;
m class index of maximum probability per video: $\mathbf{K}_{flatten} \in R^{b\times l\cdot m} \leftarrow Flatten\left(\mathbf{K} \in R^{b\times l\times m}\right)$ $\mathbf{K}_{video\_max} \in R^{b\times m} \leftarrow top\_m(\mathbf{K}_{flatten} \in R^{b\times l\cdot m})$;
Loss calculation:
for i *in* $range(b)$ **do**
  for j *in* $range(l)$ **do**
    $\mathrm{a} = \mathrm{set}\left(\mathbf{K}_{frame\_max}[i, j, \ldots] \in \mathrm{R}^{1\times m}\right)$;
    $\mathrm{d} = \mathrm{set}\left(\mathbf{K}_{video\_max}[i, \ldots] \in \mathrm{R}^{1\times m}\right)$;
    $\mathrm{loss}_{\mathrm{class}} += \left[1 - \frac{\mathrm{inter\,sec\,tion(a,d)}}{m}\right]$;
$\mathrm{loss}_{\mathrm{class}} = \mathrm{loss}_{\mathrm{class}} / (b \times l)$.

2) Video Frame Feature Coding Network: The pre-trained CLIP(ViT-B/32) [29] image feature encoder is transferred to the video frame feature encoder. Each video frame is encoded using a pre-trained CLIP image encoder. The pre-trained frame encoder first uses ViT [8] to segment each frame into non-overlapping blocks, and then linearly maps each

block feature to a 1D tensor. Each frame can be divided into k non-overlapping blocks $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{k-1}, \mathbf{p}_k\}$. The tensors of k different blocks of each frame are input into the Transformer model, and the feature representation $\mathbf{o} \in \mathrm{R}^{1\times e}$ of the frame is finally obtained. Where, e represents the dimension embedded in each frame. The feature of each video sampled l frame is represented as $\{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{l-1}, \mathbf{o}_l\}$. The video feature represented by batch is $\mathbf{F} \in \mathrm{R}^{b\times l\times e}$. We adopt the average pooling strategy to pool l frames $\mathbf{f} = \text{mean\_pooling}(\{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{l-1}, \mathbf{o}_l\}) \in \mathrm{R}^{1\times e}$ of each video, and finally generate the video feature representation $\mathbf{M} = \{\mathbf{f}_i\}_{i=1}^{b} \in \mathrm{R}^{b\times e}$.

3) Video Internal Frame Aggregation Loss Module: Batch feature $\mathbf{V}$ Generates feature $\mathbf{F} \in \mathrm{R}^{b\times l\times e}$ through video coding backbone network. Average pooling of $\mathbf{F}$ produces $\mathbf{M} \in \mathrm{R}^{b\times e}$. Feature $\mathbf{F}$ and feature $\mathbf{M}$ are used to calculate the video internal frame aggregation loss, and the pseudocode is shown in Algorithm 2.

Algorithm 2 Video Internal Frame Aggregation Loss Calculation

input : $\mathbf{F} \in \mathrm{R}^{b\times l\times e}$ and $\mathbf{M} \in \mathrm{R}^{b\times e}$
output: $\text{loss}_{\text{internal}}$

Loss calculation:
for *i in range*(b) **do**
  for *j in range*(l) **do**
    $\mathbf{z} = \mathrm{F}[i, j, \ldots] \in \mathrm{R}^{1\times e}$;
    $\mathbf{h} = \mathbf{M}[i, \ldots] \in \mathrm{R}^{1\times e}$;
    $\text{loss}_{\text{internal}} += \text{cross\_entropy}(\mathbf{z}, \mathbf{h})$;
$\text{loss}_{\text{internal}} = \text{loss}_{\text{internal}} / (b \times l)$.

B. Text Feature Processing Network Branch

Each text feature can be represented as $\mathbf{t} = \{\mathrm{w}_i\}_{i=1}^{u}$. Where u represents the number of words in a text. The model batch reads the corresponding text data $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^{b} \in \mathrm{R}^{b\times u}$ of the video. A text encoder $\mathrm{g}(\mathbf{T})$ is pre-trained with CLIP to generate text feature representation $\mathbf{C} = \mathrm{g}(\mathbf{T}) = \{\mathbf{c}_i\}_{i=1}^{b} \in \mathrm{R}^{b\times e}$.

C. Similarity Measure

We calculate the cosine similarity between video and text. The measurement formula is as follows:

$$\text{sim}\left(\mathbf{c}_i, \mathbf{f}_j\right) = \frac{\mathbf{c}_i \mathbf{f}_j}{\|\mathbf{c}_i\| \|\mathbf{f}_j\|} \tag{1}$$

and the video retrieval text loss and text retrieval video loss are obtained by measuring the similarity between video and text:

$$\text{loss}_{\mathrm{v}\to\mathrm{t}} = -\frac{1}{b}\sum_{i=1}^{b} \log \frac{\exp\left(\text{sim}\left(\mathbf{c}_i, \mathbf{f}_i\right)\right)}{\sum_{j=1}^{b} \exp\left(\text{sim}\left(\mathbf{c}_j, \mathbf{f}_i\right)\right)} \tag{2}$$

$$\text{loss}_{\mathrm{t}\to\mathrm{v}} = -\frac{1}{b}\sum_{i=1}^{b} \log \frac{\exp\left(\text{sim}\left(\mathbf{c}_i, \mathbf{f}_i\right)\right)}{\sum_{j=1}^{b} \exp\left(\text{sim}\left(\mathbf{c}_i, \mathbf{f}_j\right)\right)} \tag{3}$$

The retrieval loss is defined as:

$$\text{loss}_{\text{retrieval}} = \text{loss}_{\mathrm{v}\to\mathrm{t}} + \text{loss}_{\mathrm{t}\to\mathrm{v}} \tag{4}$$

and the total loss of video and text cross-modal retrieval is as follows:

$$\text{loss} = \alpha \text{loss}_{\text{retrieval}} + \beta\left(\text{loss}_{\text{class}} + \text{loss}_{\text{internal}}\right) \tag{5}$$

where, α and β are hyper parameters.

IV. Experiments

A. Datasets

MSVD dataset [30] contains a total of 1970 videos, each lasting about 62 seconds, and each video has 40 related English descriptions. The dataset is divided into 1200 videos in the training set, 100 videos in the verification set and 670 videos in the test set.

MSR-VTT dataset [31] contains a total of 10,000 videos and 200,000 concepts, each lasting approximately 20 seconds. In our experiments, the division of dataset is the same as HowTo100M [7]. The experiments use 7,000 videos for training and 1,000 videos for testing, and the division of test set is the same as that of JSFusion [4].

DiDeMo dataset [32] contains a total of 10,000 videos and 40,000 sentences. According to [4], [9], we verify video to sentence retrieval, and we associate multiple text descriptions of the video into a paragraph request.

B. Evaluation Indicator

We use the most commonly used metrics for cross-modal retrieval, namely Recall at Rank n (R@n), Median Rank (MdR) and Mean Rank (MnR). R@n (the bigger the better) represents the percentage of matches in the first n items of the retrieval results. MdR(a smaller value is better) represents the median number of correct search results ranked in the search results. MnR (the smaller the better) represents the average of the correct search result rankings in the search results.

C. Baseline Methods

MUL-cues [33] uses multiple modal cues to learn joint embedded representations. A fusion strategy is proposed to combine embedding learning with multiple modal features.

CE [6] uses the collaborative expert method to embed multiple modal features contained in the video and finally obtains the video feature representation. The method uses pre-trained extractors (experts) to extract features of multiple modalities from the raw video and then fuses the features of multiple modalities.

Supact-set [34] proposes a noise contrast learning method to improve the similarity of related video and text features. The method uses a generation model to gather related video samples together, and the title of each video sample needs to be reconstructed by a weighted combination method.

TeachText [35] proposes a generalized distillation method using information from multiple text encoders as a complementary cue to provide an enhanced supervisory signal for the model.

Frozen [36] is an end-to-end training model, which simultaneously uses large-scale image and video subtitle data to train the network. The model is an extension of the recent ViT model and can be trained directly on image and text datasets. On video and text datasets, the static state of video is treated as a static image for training.

CLIP [9] uses a contrastive language image pre-training model to construct a visual and textual semantic space, and then directly projects the video and text into the learned visual and textual semantic space for learning, which alleviates the difficulty of video annotation.

HowTo100M [7] uses video captions to generate text representations directly, without manual annotation of the video dataset. Finally, the video and the generated text data are learned to embed representations.

Less-more [10] proposes an end-to-end training model for sparse sampling of video clips. The method divides the whole video into several short video clips, and each video clip is encoded by a video encoder. The model uses sparse sampling method to greatly reduce the size of the model's training data.

CMHM [5] utilizes the hierarchical sequence embedding model to embed sequence data of different modalities into the hierarchical semantic space. The model is trained on large-scale video and paragraph retrieval datasets and applied to downstream tasks.

W2VV [37] uses a deep neural network architecture that can predict visual feature representations from text. In this model, text embedded features are converted into depth visual features by multi-layer perceptron.

VSE++ [38] uses data enhancement and fine-tuning methods to innovate on the basis of common loss functions of existing multi-modality and achieve better performance.

TCE [39] designs a tree-enhanced query encoder and a time-sensitive video encoder. The two encoders can not only extract the spatial information of video features, but also mine the temporal information of video features.

HGR [40] builds a hierarchical graph inference model, which can match video and text features from global features and local features, and deduce the relationship between visual features and text features from different levels.

DE [26] uses the method of hybrid space learning and adopts symmetrical model design, thus extracting coarse-grained features and fine-grained features of video and text. The model uses LSTM network to extract temporal information of video features.

FSE [5] embeds video and text features to different levels, and embeds sequence data of different modalities into the hierarchical semantic space, so as to mine the implicit and explicit information corresponding to the two modalities.

VSE [33] learns the method of joint representation to establish the semantic relationship between video and text features, and uses the multi-modal cues in video to mine the visual semantic information contained in video.

D. Comparative Experiments

Comparative experiments are performed on three commonly used benchmark datasets MSVD, MSR-VTT and DiDeMo, and our proposed approach has achieved good results. The experimental results on the MSVD dataset are shown in TABLE I, from which it can be seen that our method is superior to the comparative research methods. The MSR-VTT dataset is the most commonly used dataset for video and text cross-modal retrieval performance evaluation, and the comparative experimental results on this dataset are shown in TABLE II. TABLE III shows a comparative experiments on the DiDeMo dataset. Some of the paper's methods could not be replicated, resulting in some data not available.

TABLE I
COMPARATIVE RESULTS ON MSVD DATASET

Method	t→v					v→t				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
VSE(2018)	12.3	30.1	42.3	14.0	-	34.7	59.9	70.0	3.0	-
MUL-cues(2018)	20.3	47.8	61.1	6.0	28.3	31.5	51	61.5	5.0	41.7
CE(2019)	19.8	49	63.8	6.0	23.1	-	-	-	-	-
Supact-set(2021)	23	52.8	65.8	5.0	-	27.3	50.7	60.8	5.0	-
TeachText(2021)	25.7	57.3	71.5	4.0	-	-	-	-	-	-
Frozen(2020)	33.7	64.7	76.3	3.0	-	-	-	-	-	-
CLIP(2021)	37	64.1	73.8	3.0	-	59.9	85.2	90.7	1.0	-
ours	**45.4**	**75.6**	**84.6**	**2.0**	**10.3**	**64.9**	**88.7**	**94.4**	**1.0**	**3.4**

TABLE II
COMPARATIVE RESULTS ON MSR-VTT DATASET

Method	t→v					v→t				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
W2VV(2018)	2.7	12.5	17.3	83	-	17.3	42	53.5	9	-
VSE++(2018)	17	40.9	52	10	-	18.1	40.4	52.1	9	-
MUL-cues(2018)	12.25	32.1	42.4	16	134	7	20.9	29.7	38	213.8
HowTo100M(2019)	14.9	40.2	52.8	9	-	-	-	-	-	-
TCE(2020)	17	44.7	58.3	7	-	15.1	43.3	58.2	7	-
HGR(2020)	22.9	50.2	63.6	5	-	20	48.3	60.9	6	-
Less-more(2021)	11.6	33.9	45.8	13	-	-	-	-	-	-
TeachText(2021)	31.2	53.7	64.2	3	-	27.2	51.7	62.6	5	-
DE(2021)	23	50.6	62.5	5	-	25.1	52.1	64.6	5	-
ours	**42.2**	**67.9**	**78.7**	**2**	**16.7**	**40.2**	**68.7**	**79.3**	**2.0**	**12.7**

TABLE III
COMPARATIVE RESULTS ON DIDEMO DATASET

Method	t→v					v→t				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
CMHM(2018)	30.2	60.5	-	3.3	-	30.1	59.2	-	3.0	-
FSE(2018)	14.6	36.8	80.5	13.0	-	13.6	34.3	78.8	12	-
CE(2019)	16.1	41.1	82.7	8.3	43.7	15.6	40.9	82.2	8.2	42.4
Less-more(2021)	20.4	48	60.8	6.0	-	-	-	-	-	-
Frozen(2021)	34.6	65	74.7	3.0	-	-	-	-	-	-
ours	**38.3**	**66.3**	**76.8**	**2.0**	**22.1**	**39.2**	**65**	**75.5**	**3.0**	**16.3**

E. Implementation Details

The model implements the Pytorch, and the server is configured with NVIDIA Titan RTX24GB. The program is trained on four graphics cards in a distributed manner with the learning rate and the batch size being respectively set to 3E-5 and 128. The decay rate of learning rate is 0.8. α is set to 0.7 and β is set to 0.3. The number of each video sampled frame is 12. The maximum number of words per sentence is set to 32. The batch size of data during validation is set to 20.

V. ABLATION EXPERIMENTS

In order to verify the effectiveness of our proposed model, we study the ablation of video frame classification-aggregation

TABLE IV
RESULTS ON MSVD DATASET UNDER DIFFERENT CONDITIONS

Model name	t→v			v→t		
	R@1	R@5	R@10	R@1	R@5	R@10
Base	42.8	72	81.7	57.8	83.4	88
Base+C	44.9	74.7	83.8	63.9	88	93
Base+C+I	**45.4**	**75.7**	**84.6**	**64.9**	**88.7**	**94.4**

TABLE V
THE RESULTS UNDER DIFFERENT VALUE OF m

Number of objects(m)	t→v			v→t		
	R@1	R@5	R@10	R@1	R@5	R@10
5	44.9	75.0	84.1	64.2	87.8	93.6
10	**45.4**	**75.7**	**84.6**	**64.9**	**88.7**	**94.4**
15	45.2	74.9	84.2	64.4	88.2	94.0
20	45.0	74.9	84.1	64.3	88.0	93.8

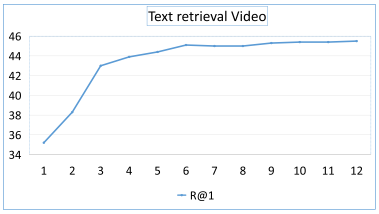
network and video internal frame aggregation loss module. Firstly, the video encoder and text encoder of pre-trained CLIP are used as the basic models to verify the performance of the video and text cross-modal retrieval model. This base model is denoted as "Base". Then we add the video frame classification-aggregation network and denote it as "Base+C". Finally, the video internal frame aggregation loss module is added which is denoted as "Base+C+I". The results on MSVD dataset are reported in TABLE IV.

It can be observed from TABLE IV that, each part influences the video and text cross-modal retrieval. In order to verify the influence of video frame classification-aggregation network on the overall performance of the model, we conducted ablation experiments on the number of categories m with high probability of occurrence in each sampled frame. The results are shown in TABLE V. It can be seen from TABLE V that when the quantity of m is 10, the performance of this model reaches the best. When m is small, it indicates that the classes of important objects in each video sampled frame are reduced, which leads to the reduction of important object features contained in the whole video feature. When m is large, the performance will also decline, which indicates that excessive object features in the video will lead to information redundancy, but adversely affect the retrieval performance. This redundant information is equivalent to noise. Because the number of important objects contained in the text describing the video is limited, if the number of important objects contained in the video feature extraction is much larger than that contained in the text, it may cause the existence of information redundancy in the video feature.

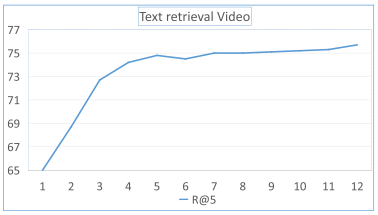
In the video internal frame aggregation loss module, we preliminarily explore the temporal information of different sampled frames in the video. We input feature $\mathbf{F}$ into LSTM network to produce feature $\mathbf{F}'$ with temporal information, and by average pooling $\mathbf{F}'$ to produce feature $\mathbf{M}$. TABLE VI shows the results after processing the temporal features of the sampled frames. Unfortunately, it didn't have the effect as expected. The reason is that the LSTM network introduced a large number of parameters when it is added to establish temporal information between sampled frames. If the training time is not enough, these parameters can not be optimized. If the training time of cross-modal retrieval network increases, the pre-trained model will be over-fitted, which will gradually reduce the performance. In the future we will continue to explore how to mine temporal information in video sampled frames.

TABLE VI
EVALUATION ON THE TEMPORAL INFORMATION

Model name	t→v			v→t		
	R@1	R@5	R@10	R@1	R@5	R@10
Base+C+I	**45.4**	**75.7**	**84.6**	**64.9**	**88.7**	**94.4**
Base+C+I+LSTM	41.0	71.5	79.4	57.5	83.1	88.0



(a) Text retrieval video R@1 accuracy.



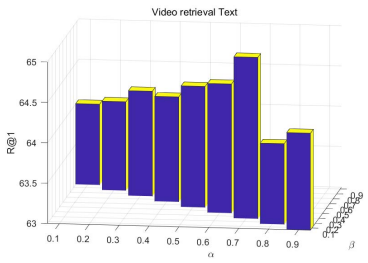
(b) Text retrieval video R@5 accuracy.

Fig. 5. Results under different number of sampled frames from each video.

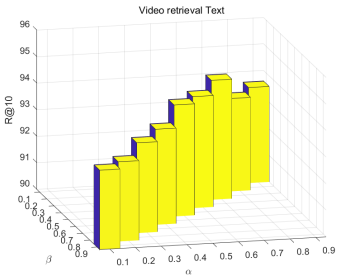
We also evaluate the influence of the number of sampled frames on the retrieval accuracy, and report the results on MSVD dataset in Fig. 5. It can be found that, with the increase of the number of sampled frames, the increase rate of the accuracy gradually decreases. As the number of video sampled frames increases, the whole video important objects form a collection, and the number of elements in the collection is gradually fixed. Important object features in this object collection should be more prominent, therefore, mining the important classification features contained in each frame is of great help to cross-modal retrieval. Since those unimportant features may interfere with the retrieval, it is necessary for us to use video frame classification-aggregation network.

It can be observed from Fig. 5 that when the number of sampled frames is 10, the model can almost achieve the optimal result. In the actual experiments, it can be set appropriately according to the hardware device. The parameters of the model increase with the increase of the number of sampled frames, which may increase the requirements on computer hardware. In our experiments, 12 frames are sampled.

There are two group losses during model training. The first group is the video and text retrieval $\text{loss}_{\text{retrieval}}$, and the second group is the sum of video frame classification-aggregation $\text{loss}_{\text{class}}$ and video internal frame aggregation $\text{loss}_{\text{internal}}$, which



(a) Video retrieval text R@1 accuracy.



(b) Video retrieval text R@10 accuracy.

Fig. 6. Results under different α and β values.

(a) Change in $\text{loss}_{\text{retrieval}}$.

(b) Chang in loss_2.

Fig. 7. Changes in $\text{loss}_{\text{retrieval}}$ and loss_2 values during training.

is expressed as $\text{loss}_2 = \text{loss}_{\text{class}} + \text{loss}_{\text{internal}}$. We adopt two hyperparameters (α, β) to combine the two group losses, and analyze the influence of these two hyperparameters on the overall performance of the model on MSVD dataset. The sum of α and β equals 1, and the results are shown in Fig. 6. We can see from Fig. 6 that, the model achieves the optimal results when α=0.7 and β=0.3. Among them, the change of loss value during training is shown in Fig. 7.

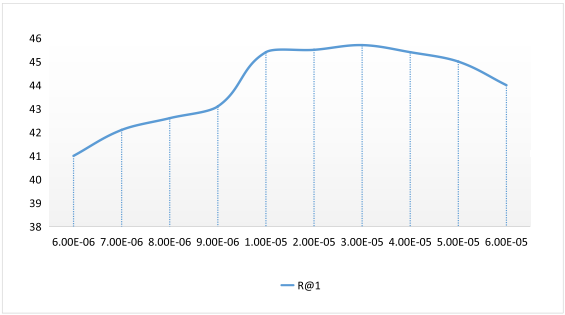


Fig. 8. Recall rate (R@1) of text retrieval video changes with learning rate.

The learning rate is an important parameter in neural network learning, which influences the convergence rate of the network training. We have done a large number of experiments on MSVD dataset, and find that the optimal learning rate is 3E-5. The experimental results are shown in Fig. 8.

VI. Conclusion

Sparse sampling of video frames is used to extract video features, which enables the image encoder to encode video frame features successfully. We apply video frame sparse sampling technique to video feature extraction so as to further explore the relationship between the sampled frames. By designing the video frame classification-aggregation network, the features of important objects contained in the video can be highlighted. This enables the visual focus the whole video on important objects, thus facilitating the alignment of video modal features and text modal features on the important object features, and improving the similarity between text features and video features. The construction of video internal frame loss module further reduces the difference between features of different frames after the video frame is encoded. The proposed model not only reduces the semantic difference between different modalities to make the distribution of different modal features tend to be consistent through network training, but also reduces the difference between video modal frames to make the video internal frame features should tend to be consistent. In the future, we need to further study the influence of sampled video frames on retrieval performance. Sparse sampling is also adopted in video frame sampling, since there is a temporal problem between different sampling frames. The number of video sampling frames should also be different for different retrieval tasks, since excessive number may lead to information redundancy of video features. How to determine the sample number and sampling rate for a specific retrieval task needs further research in our future work.

References

[1] L. Zhu, H. Cui, Z. Cheng, J. Li, and Z. Zhang, "Dual-level semantic transfer deep hashing for efficient social image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1478–1489, Apr. 2021.

[2] J. Wu, C. Wu, J. Lu, L. Wang, and X. Cui, "Region reinforcement network with topic constraint for image-text matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 388–397, Jan. 2022.

[3] A. Miech, I. Laptev, and J. Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," 2018, *arXiv:1804.02516*.

[4] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11211. Berlin, Germany: Springer, 2018, pp. 487–503.

[5] B. Zhang, H. Hu, and F. Sha, "Cross-modal and hierarchical modeling of video and text," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11217, 2018, pp. 385–401.

[6] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," in *Proc. 30th Brit. Mach. Vis. Conf. (BMVC)*, Cardiff, U.K., Sep. 2019, p. 279.

[7] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2630–2640.

[8] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[9] J. A. Portillo-Quintero, J. C. Ortiz-Bayliss, and H. Terashima-Marín, "A straightforward framework for video retrieval using CLIP," in *Proc. Mex. Conf. Pattern Recognit.*, vol. 12725, 2021, pp. 3–12.

[10] J. Lei *et al.*, "Less is more: ClipBERT for video-and-Language learning via sparse sampling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7331–7341.

[11] W. Jin, Z. Zhao, P. Zhang, J. Zhu, X. He, and Y. Zhuang, "Hierarchical cross-modal graph consistency learning for video-text retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1114–1124.

[12] Z. Jiang, V. Rozgic, and S. Adali, "Learning spatiotemporal features for infrared action recognition with 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 309–317.

[13] X. He, Y. Peng, and L. Xie, "A new benchmark and approach for fine-grained cross-media retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1740–1748.

[14] X. Yu, T. Chen, Y. Yang, M. Mugo, and Z. Wang, "Cross-modal person search: A coarse-to-fine framework using bi-directional text-image matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1799–1804.

[15] Z. Wang *et al.*, "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5763–5772.

[16] Y. Peng and J. Chi, "Unsupervised cross-media retrieval using domain adaptation with scene graph," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4368–4379, Nov. 2020.

[17] S. Qian, D. Xue, H. Zhang, Q. Fang, and C. Xu, "Dual adversarial graph neural networks for multi-label cross-modal retrieval," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2021, pp. 2440–2448.

[18] X. Dong, L. Liu, L. Zhu, L. Nie, and H. Zhang, "Adversarial graph convolutional network for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1634–1645, Mar. 2022.

[19] X. Dong, H. Zhang, L. Zhu, L. Nie, and L. Liu, "Hierarchical feature aggregation based on transformer for image-text matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6437–6447, Sep. 2022.

[20] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11208. Berlin, Germany: Springer, 2018, pp. 212–228.

[21] Z. Hu, Y. Luo, J. Lin, Y. Yan, and J. Chen, "Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, Macao, China, Aug. 2019, pp. 789–795.

[22] Y. Peng, J. Qi, and Y. Zhuo, "MAVA: Multi-level adaptive visual-textual alignment by cross-media bi-attention mechanism," *IEEE Trans. Image Process.*, vol. 29, pp. 2728–2741, 2020.

[23] Y. Hu, M. Liu, X. Su, Z. Gao, and L. Nie, "Video moment localization via deep cross-modal hashing," *IEEE Trans. Image Process.*, vol. 30, pp. 4667–4677, 2021.

[24] M. Qi, J. Qin, Y. Yang, Y. Wang, and J. Luo, "Semantics-aware spatial–temporal binaries for cross-modal video retrieval," *IEEE Trans. Image Process.*, vol. 30, pp. 2989–3004, 2021.

[25] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "COOT: Cooperative hierarchical transformer for video-text representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 22605–22618.

[26] J. Dong *et al.*, "Dual encoding for video retrieval by text," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4065–4080, Aug. 2022.

[27] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12349, 2020, pp. 214–229.

[28] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 6105–6114.

[29] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Virtual Event, Austria, 2021, pp. 1–22.

[30] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 190–200.

[31] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.

[32] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5804–5813.

[33] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2018, pp. 19–27.

[34] M. Patrick *et al.*, "Support-set bottlenecks for video-text representation learning," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, Virtual Event, Austria, 2021, pp. 1–18.

[35] I. Croitoru *et al.*, "TeachText: CrossModal generalized distillation for text-video retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11563–11573.

[36] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1708–1718.

[37] J. Dong, X. Li, and C. G. M. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3377–3388, Dec. 2018.

[38] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Newcastle, U.K. Sep. 2018, p. 12.

[39] X. Yang, J. Dong, Y. Cao, X. Wang, M. Wang, and T.-S. Chua, "Tree-augmented cross-modal encoding for complex-query video retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Virtual Event, China, Jul. 2020, pp. 1339–1348.

[40] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 10635–10644.

Ming Jin received the bachelor's degree from Shaoyang University in 2020. He is currently a Graduate Student at the School of Information Science and Engineering, Shandong Normal University, China. He has been under the supervision of Prof. Huaxiang Zhang at Shandong Normal University since 2020. His research interests include large-scale multimedia content analysis and retrieval.

Huaxiang Zhang received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2004. He was an Associate Professor with the Department of Computer Science, Shandong Normal University, Jinan, China, from 2004 to 2005. He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University, together with the School of Information Science and Electrical Engineering, Shandong Jiaotong University. He has authored over 200 journals and conference papers and has been granted 31 invention patents, and is supported by the Program of Taishan Scholar. His current research interests include machine learning, pattern recognition, evolutionary computation, and multimedia analysis.

Lei Zhu (Senior Member, IEEE) received the B.S. degree from the Wuhan University of Technology in 2009, and the Ph.D. degree from the Huazhong University of Science and Technology in 2015. He was a Research Fellow under the supervision of Prof. Heng Tao Shen at the University of Queensland from 2016 to 2017, and Dr. Jialie Shen at Singapore Management University from 2015 to 2016. He is currently a Full Professor with the School of Information Science and Engineering, Shandong Normal University, China. His research interests include large-scale multimedia content analysis and retrieval.

Jiande Sun received the Ph.D. degree in communication and information system from Shandong University, Jinan, China, in 2005. From 2008 to 2009, he was a Visiting Researcher with the Institute of Telecommunications System, Technical University of Berlin, Berlin, Germany. From 2010 to 2012, he was a Post-Doctoral Researcher with the Institute of Digital Media, Peking University, Beijing, China, and also with the State Key Laboratory of Digital-Media Technology, Hisense Group, Qingdao, China. From 2014 to 2015, he was a DAAD Visiting Researcher with the Technical University of Berlin and the University of Konstanz, Konstanz, Germany. From 2015 to 2016, he was a Visiting Researcher with the School of Computer Science, Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University. He has authored or coauthored more than 60 journals and conference papers and is the coauthor of two books. His current research interests include multimedia content analysis, video hashing, gaze tracking, image/video watermarking, and 2-D-to-3-D conversion.

Li Liu received the Ph.D. degree from Shandong University in 2007. She is currently a Professor and a Master's Supervisor with the School of Information Science and Engineering, Shandong Normal University. She has authored more than 40 academic articles. Her research interests include computer graphics.