BiC-Net: Learning Efficient Spatio-temporal Relation for Text-Video Retrieval

NING HAN, School of Computer Science, Xiangtan University, China
YAWEN ZENG, Bytedance AI Lab, China
CHUHAO SHI, GUANGYI XIAO, and HAO CHEN, College of Computer Science and Electronic Engineering, Hunan University, China
JINGJING CHEN, School of Computer Sciences, Fudan University, China

The task of text-video retrieval aims to understand the correspondence between language and vision and has gained increasing attention in recent years. Recent works have demonstrated the superiority of local spatio-temporal relation learning with graph-based models. However, most existing graph-based models are handcrafted and depend heavily on expert knowledge and empirical feedback, which may be unable to mine the high-level fine-grained visual relations effectively. These limitations result in their inability to distinguish videos with the same visual components but different relations. To solve this problem, we propose a novel cross-modal retrieval framework, Bi-Branch Complementary Network (BiC-Net), which modifies Transformer architecture to effectively bridge text-video modalities in a complementary manner via combining local spatio-temporal relation and global temporal information. Specifically, local video representations are encoded using multiple Transformer blocks and additional residual blocks to learn fine-grained spatio-temporal relations and long-term temporal dependency, calling the module a Fine-grained Spatio-temporal Transformer (FST). Global video representations are encoded using a multi-layer Transformer block to learn global temporal features. Finally, we align the spatio-temporal relation and global temporal features with the text feature on two embedding spaces for cross-modal text-video retrieval. Extensive experiments are conducted on MSR-VTT, MSVD, and YouCook2 datasets. The results demonstrate the effectiveness of our proposed model. Our code is public at https://github.com/lionel-hing/BiC-Net.

CCS Concepts: • Information systems → Multimedia and multimodal retrieval;

Additional Key Words and Phrases: Text-video retrieval, spatio-temporal relation, bi-branch complementary network

This work is supported by the National Natural Science Foundation of China (grant no. 62272156) and the Shanghai Pujiang Program (grant no. 20PJ1401900).

Authors’ addresses: N. Han, School of Computer Science, Xiangtan University, Yuhu District, Xiangtan, China, 411105; e-mail: ninghan@xtu.edu.cn; Y. Zeng, Bytedance AI Lab, 43 North Third Ring West Road, Beijing, China, 100098; e-mail: yawenzeng11@gmail.com; C. Shi, G. Xiao (Corresponding author), and H. Chen, College of Computer Science and Electronic Engineering, Hunan University, 116 Lu Shan South Road, Yuelu District, Changsha, China, 410082; e-mails: sch8288@hnu.edu.cn, guangyi.xiao@gmail.com, chenhao@hnu.edu.cn; J. Chen, School of Computer Sciences, Fudan University, 20 Handan Road, Yangpu District, Shanghai, China, 200433; e-mail: chenjingjing@fudan.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1551-6857/2023/12-ART86 $15.00

https://doi.org/10.1145/3627103

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 20, No. 3, Article 86. Publication date: December 2023.
1 INTRODUCTION

Recent years have witnessed an exponential growth of multimedia data \([8, 63, 66, 73]\) (e.g., video, image, and text), which has increased the demand for effectively retrieving relevant data from another modality when given a query of one modality. As one of these challenging tasks, text-video retrieval aims to retrieve video given a text query, which requires measuring the semantic similarity between a sentence and a video. Video data are distinct from images due to the temporal dependencies among frames and the additional dynamic relationships among objects, resulting in the inability of existing video retrieval techniques to distinguish videos with the same visual components but with different relations. Figure 1 shows such an example. Given the text query "A woman breaks the two bacon slices into pieces and lays them on the tomatoes", the existing retrieval systems are likely to consider both (a) and (b) as positive examples since both contain the same motion ("laying") and objects ("bacon slices", "tomatoes") as the text query. However, example (b) is indeed a false positive, as it presents "a woman lays tomatoes on the green leaves" (with bacon slices on the kitchen table). This example suggests that ignoring the visual relations (i.e., object relations) presented in videos could lead to inaccurate retrieval results. Therefore, capturing high-level spatio-temporal visual relations in videos is crucial to distinguish similar videos.

This article investigates the problem of cross-modal text-video retrieval. In the literature, many efforts have been devoted to learning better video representations to improve text-video retrieval performance. Based on the granularity of feature representations, existing works can be roughly categorized into global and local feature-based methods. Global feature-based methods typically use global representations to represent entire videos and sentences, which usually lose part of this temporal information and local details. Such approaches work well in a simple cross-modal retrieval scenario, in which only a single object is presented in the video or text query. For more realistic cases involving complex natural scenes, the performance of these methods is usually unsatisfactory. In contrast, local feature-based methods pay attention to local details and perform matching by detecting objects in videos and texts. With local region modeling, the performance of text-video retrieval has been significantly improved. Nevertheless, the existing efforts can only capture simple visual relations with a graph convolutional network (GCN) \([15, 46]\) or utilize an attention mechanism \([9, 59]\) as a cross-modal interaction module to delve into high-level correspondences. As GCN-based video modeling is handcrafted, it depends heavily on expert knowledge and empirical feedback, which may not be able to effectively mine and model the high-level fine-grained visual relations. Attention-based models, on the other hand, selectively align the key information presented in different modalities. As attention-based methods also ignore fine-grained visual relations, the performance of these methods is still unsatisfactory. Thus, novel modeling solutions are eagerly awaited.

To further improve the performance of text-video retrieval, this article studies this problem from the perspective of spatio-temporal relation modeling for videos. Generally, there are two major obstacles in modeling the spatio-temporal relation. First, videos contain diverse spatial and temporal information within variations in motion and richer information in local visual details. These objects and interactions increase the difficulties in capturing high-level fine-grained visual content. Second, local relation modeling captures considerable fragmented information, which
BiC-Net: Learning Efficient Spatio-Temporal Relation for Text-Video Retrieval

**Textual Query:** A **woman breaks** the two **bacon slices** into pieces and **lays** them on the **tomatoes**.

![Image](image_url)

**Fig. 1.** An example of text-to-video retrieval. Given a textual query, a common pipeline with fine-grained [46] or global semantical visual features [36] will return two videos with the same compositions. Our BiC-Net can accurately model the temporal and spatial interactions between objects (e.g., ‘bacon’ lays on ‘tomatoes’) to filter out false-positives without correct interactions.

will overlook contextual information. Therefore, the way to comprehensively capture multi-granularity visual information to represent videos from complementary spatial and temporal perspectives is of great importance.

To address the aforementioned problems, we propose a novel Bi-Branch Complementary Network (BiC-Net), which modifies transformer architecture to effectively bridge text-video modalities in a complementary manner via combining local spatio-temporal relations and global temporal information. We present an overview of BiC-Net in Figure 2. Specifically, for videos, our BiC-Net attempts to extract two perspectives of features – global temporal features and local relation features. At the global temporal level, we directly adopt the widely used 2D and 3D-CNN. For local relational features, we use pre-trained Faster R-CNN [43] to extract regional features (i.e., features of bounding boxes). Then, a fine-grained spatio-temporal transformer is employed for learning fine-grained spatio-temporal relations and long-term temporal dependency. This module separately captures local spatial and long-term temporal relations among local spatial relations. In addition, a multi-layer transformer block is applied for learning global temporal features. To cover different levels of semantics, we align the global temporal and local relation features with the text feature on two embedding spaces. Lastly, the similarity between videos and texts is measured in both embedding spaces and then summed to obtain the final similarity score. In this way, the global temporal information and local relation information in a video can be utilized for cross-modal text-video retrieval comprehensively.

Our contributions are summarized as follows.

- We incorporate feature-split with a bi-branch framework called BiC-Net to capture local relations and global temporal features comprehensively, which aligns the global temporal and local relation features with the text feature on two embedding spaces for cross-modal text-video retrieval.
Fig. 2. The overall framework of our proposed BiC-Net. First, we extract local relational and global visual features for videos. The local relations are represented by local regional features using a fine-grained spatio-temporal transformer. The global video features are represented by 2D-CNN and 3D-CNN features via a multi-layer temporal transformer. Then, we extract textual features by BERT. Finally, both video and relational features are leveraged to align with textual features on two embedding spaces for cross-modal text-video retrieval. Among them, T-Block denotes the transformer block.

- We first introduce a simple and effective fine-grained spatio-temporal transformer to learn fine-grained spatio-temporal relations and long-term temporal dependency and a multi-layer temporal transformer to further explore global temporal information for global temporal features. In this way, the bi-branch information in video and text can accurately capture cross-modal semantic alignment in a cooperative and complementary manner.
- We conduct extensive experiments on three standard benchmarks and verify the effectiveness of our proposed method by showing that BiC-Net can achieve SOTA performance (86.7% on the MSR-VTT 1k-A test set) under similar conditions.

The rest of this article is organized as follows. In Section 2, we briefly present a review of related work. In Section 3, we describe our proposed BiC-Net model. In Section 4, we provide implementation details and experimental results. In Section 5, we conclude our article.

2 RELATED WORK

2.1 Text-Video Retrieval

According to the granularity of feature representations, we roughly divide existing works into two groups: global feature–based methods and local relation feature–based methods.

**Global feature–based methods** [13, 36, 62, 64, 71] extract global feature representations of videos and texts and then learn a joint embedding space where visual and textual similarity is measured. For the video representation, they adopt 2D/3D CNN models to extract frame features and aggregate frame features by average-pooling [11, 28, 31] or max-pooling [35, 36]. For the video representation, they focused only on leveraging the global feature of the video. For instance, Dong et al. [12, 13] employ three levels, i.e., global, temporal, and local, to encode videos and texts and learn a hybrid common space for video-text similarity measurement. Miech et al. [36] adopt 2D and 3D CNN models to extract frame features and only use max-pooling to obtain global video representation. Yang et al. [62] present a latent semantic tree to encode the text and used a multi-head self-attention mechanism to obtain the temporal-attentive video representation.

**Local feature–based methods** [9, 15, 19, 20, 57, 59] use local semantic information from language or video for better text-video alignment from different aspects and then perform text-video
retrieval tasks. Wray et al. [57] disentangle action phrases into verbs and nouns for fine-grained video retrieval. The graph-based approaches [15, 20, 46] construct different semantic correlation graphs for videos and learn fine-grained semantic relations for text-video retrieval. Some works [9, 19, 59] also propose fine-grained alignment models that decompose text and video into multiple levels and align text with video at multiple levels for text-video matching. Similarly, studies in the literature [30, 51–53] introduce interactive alignment networks aimed at enhancing model accuracy. For instance, Liu et al. [30] introduce EARN, an entity-enhanced adaptive reconstruction network for weakly supervised relation extraction, utilizing entity-based supervision and hierarchical attention for improved handling of diverse referring expressions without requiring manual annotation of region-level ground truth. Wang et al. [52] present a semantic and relation modulation network to improve audio-visual event localization by leveraging correlations between segments and multi-scale event proposals through specialized modulating modules. The authors of [51, 53] propose a structured multi-level interaction network and a dual-path interaction network, both designed to facilitate a comprehensive understanding of visual language and intricate video comprehension.

Recently, some studies have also explored a combination of video experts (e.g., motion, audio, and speech) [16, 31, 37, 56] or pre-trained video experts [34, 38, 44] to improve the performance of cross-modal retrieval. Lately, Transformer-based works [4, 14, 27, 29, 33] have benefited from pre-training models on large-scale language-vision datasets [4, 33]. For example, Bain et al. [4] propose an end-to-end trainable model that adopts a space-time transformer encoder to train both video and image datasets flexibly. Luo et al. [33] apply the joint language-vision model of CLIP [42], pre-trained on a large-scale text-image dataset as a backbone for text-video retrieval. However, Transformer-based methods have a heavy computational burden due to computation-intensive operations and are extremely time-consuming to pre-train on large-scale datasets. Unlike these existing works, our study introduces a new fine-grained spatio-temporal transformer to learn high-level local relation features and a multi-layer transformer to further explore global temporal information for global temporal features. In this way, bi-branch information in video and text can accurately capture cross-modal semantic alignment in a cooperative and complementary manner.

2.2 Spatio-Temporal Relation Modeling in Video Understanding

For spatio-temporal relation modeling in video understanding, earlier works adopt 2D/3D CNNs to represent the core operators for spatio-temporal feature learning across downstream video tasks [32, 40, 54, 70]. However, these video representations focus on learning spatio-temporal features from the entire video and can hardly capture local spatio-temporal relation information. To understand the local relation information in the video, several efforts have demonstrated the effectiveness of incorporating local spatio-temporal relationships into video understanding in many downstream applications, such as visual relationship detection [18, 41, 47, 60], action recognition [21, 55, 68, 69], and video retrieval [20, 46, 65, 67]. For instance, Qian et al. [41] construct a spatio-temporal graph in adjacent video clips to define the relationships between objects. Wang et al. [55] abstract the video as a space-time graph for action recognition. Song et al. [46] model video as a spatio-temporal graph between object interactions for text-video retrieval. However, modeling object spatio-temporal relations in the video is still not thoroughly investigated. These studies have built visual relation graphs and adopted the GCN [26] to extract visual relation graph features. Massive graph construction and feature extraction are hand-crafted, complex, and time-consuming. Recently, the transformer [49] has shown great superiority in understanding 1, 2, and 3-dimensional signals (e.g., natural language processing and computer vision) and has strong interpretability and representation capabilities. Unlike these, our work designs a fine-grained
spatio-temporal transformer for learning the local spatio-temporal relations and further mining the object interactions. Notably, we validate in the experiments that our approach can surpass many related methods under a strict memory budget.

3 PROPOSED METHOD

As depicted in Figure 2, the overall pipeline of the proposed method consists of four modules: (1) video embedding learning, which involves extracting video global features; (2) relation embedding learning, which involves extracting local relational features in videos; (3) text embedding learning, which learns the representation for textual sentences by BERT [10]; and (4) joint embedding learning, which optimizes the correspondence between text and video features in a common space with a triple ranking loss.

3.1 Video Embedding Learning

Given a long video clip, we sample $T$ video frames from it with the same temporal duration between every two frames. For frame-level features, we first use a 2D-CNN to extract appearance features and a 3D-CNN to extract motion features. Then, we concatenate 2D and 3D features and apply a pointwise linear layer to obtain global visual features $F_g \in \mathbb{R}^{d_g}$. Finally, we feed the result to a standard multi-layer transformer block [49] and an attention-aware feature aggregation layer [17] to obtain its video embedding, which is denoted as $F_v \in \mathbb{R}^{d_v}$.

3.2 Relation Embedding Learning

In addition to having global visual features, the proposed framework learns local relation features from the video to improve the performance of cross-modal retrieval. The introduction of spatio-temporal relations among objects in the video equips the model with the ability to identify the fine-grained differences between videos with similarities. To capture the visual relations from the video, we first adopt the pre-trained Faster R-CNN [1] to detect frame-level region proposals and select the top $N$ region proposals with the highest detection confidence to represent each frame. Prior efforts [55, 58] focus on abstracting frame-level region proposals as fully connected spatio-temporal graphs and using GCN to learn relational features. However, computing all pair-wise relations across all video frames would be inefficient in creating a video as a fully connected graph. In recent years, pure transformer-based models have shown promising performance due to their strong representation capabilities. As a central piece of the transformer, self-attention comes with a flexible mechanism to deal with variable-length inputs. It can be understood as a fully connected layer where the weights are dynamically generated from pairwise relations from inputs, which conveys refreshing solutions to process visual relations.

Inspired by these pioneering efforts, to capture high-level visual relations from the video, we design a new architecture to learn the relation embeddings, named Fine-grained Spatio-temporal Transformer (FST), which exploits all the variants of transformer blocks and residual connections but composes each in the different placement of FST. In the following, the basic components used in the transformer block and the transformer block used in the FST module are presented in detail.

**Transformer Block.** The Transformer consists of multi-head self-attention (MSA), multi-layer perceptron (MLP), and layer-norm (LN). In the self-attention module, the inputs $X \in \mathbb{R}^{n \times d}$ are linearly transformed to three parts, i.e., queries $Q \in \mathbb{R}^{n \times d_k}$, keys $K \in \mathbb{R}^{n \times d_k}$, and values $V \in \mathbb{R}^{n \times d_v}$, where $n$ is the sequence length, and $d, d_k, d_v$ are the dimensions of inputs, queries (keys) and values, respectively. The scaled dot-product attention is applied on $Q, K, V$:

$$SA(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V. \quad (1)$$
With \( SA (Q, K, V) \), MSA is defined as
\[
\text{MSA} (Q, K, V) = \text{Concat} (\text{head}_1, \ldots, \text{head}_M) \ W^O,
\]
where \( \text{head}_i = \text{SA}(QW_i^Q, KW_i^K, VW_i^V) \).

Where \( QW_i^Q, KW_i^K, VW_i^V \) are projections of different heads, \( W^O \) is another mapping function. The MLP is applied between self-attention layers for feature transformation and non-linearity. It is formulated as
\[
\text{MLP} (X) = \text{GELU} (XW_1 + b_1) W_2 + b_2,
\]
where \( W_1 \in \mathbb{R}^{d \times d_m} \) and \( W_2 \in \mathbb{R}^{d_m \times d} \) are weights of the two fully connected layers, respectively, \( b_1 \in \mathbb{R}^{d_m} \) and \( b_2 \in \mathbb{R}^{d} \) are the bias terms, and GELU \([22]\) is the activation function. Layer normalization \([3]\) is a key part in transformer for stable training and faster convergence, and LN is applied over each sample \( x \in \mathbb{R}^{d} \) as follows:
\[
\text{LN} (x) = \frac{x - \mu}{\eta} \odot \gamma + \beta,
\]
where \( \mu \in \mathbb{R} \), \( \eta \in \mathbb{R} \) are the mean and standard deviation of the feature, respectively, \( \odot \) is the element-wise dot, and \( \gamma \in \mathbb{R}^{d} \), \( \beta \in \mathbb{R}^{d} \) are learnable affine transform parameters.

**FST for relation embedding learning.** We propose a fine-grained spatio-temporal transformer architecture to learn local relation information in a video. In this fine-grained spatio-temporal transformer, we have two dataflows in which one flow operates across the frame and the other processes the object proposals inside each frame. Suppose that a set of object proposals \( Y^t = \{y^t_1\}_{n=1}^N \) are in frame \( t \), where \( y^t_i \in \mathbb{R}^{d^r} \) is the feature vector of the \( n \)-th proposal and \( N \) is the top \( N \) region proposals. We view each frame tensor \( Y^t_0 \) as a sequence of object proposal embeddings:
\[
Y^t_0 = \left[ y^t_0, y^t_0, \ldots, y^t_0, y^t_0 \right].
\]

For the object proposal embeddings, to capture spatial relations among visual objects, we utilize a transformer block to explore the interaction pattern in a spatial (frame) between object proposals. Then, a residual connection is used to aggregate spatial information and original local information, formulated by
\[
Y^t_1' = Y^t_0 + \text{MSA} \left( \text{LN} \left( Y^t_{l-1} \right) \right),
\]
\[
Y^t_1'' = Y^t_1' + \text{MLP} \left( \text{LN} \left( Y^t_{l} \right) \right),
\]
\[
Y^t_1''' = Y^t_1'' + Y^t_{l-1},
\]
where \( l = 1, 2, \ldots, L \) is index of the \( l \)-th layer, and \( L \) is the total number of layers. The updated features after multi-layer transformer block are forwarded to an average pooling layer, which calculates the mean of all the proposal features and leads to a \( 1 \times d^r \) dimensions representation. All frame tensors after transformation are
\[
Z_0 = \left[ Y^L_1'', Y^L_1'', \ldots, Y^L_1'' \right].
\]

This process builds the relationship among proposals by computing interactions between any two proposals. For the frame level, we create the object proposal embedding memories to store the sequence of frame-level representations \( Z_0 \). Similar to the object proposal level processing, we use a transformer block for transforming the frame embeddings. Then, a residual connection is used to aggregate temporal information with spatial information and spatio-temporal information with original local information, respectively. Our final relation embedding is defined as
\[
Z^t_i = Z_i - 1 + \text{MSA} \left( \text{LN} \left( Z_i - 1 \right) \right),
\]
Fig. 3. Designs of FST and its variants: (a) spatial residual (Spatial-FST); (b) temporal residual (Temporal-FST); (c) spatio-temporal residual (Spatio-Temporal-FST); and (d) our FST.

\[ Z''_{l} = Z'_{l} + \text{MLP}(LN(Z'_{l})) , \]  

(11)

\[ Z'''_{l} = Z''_{l} + Z_{l-1} , \]  

(12)

\[ F_{r} = Z'''_{l} + Y_{l-1}^{t} . \]  

(13)

The temporal transformer block is used for modeling temporal relations among frame embeddings. Finally, we apply an attention-aware feature aggregation layer \[17\] to obtain the final relation embedding, denoted as \( F_{r} \in \mathbb{R}^{d_t} \).

Next, we discuss several variants for the FST, as illustrated in Figure 3. **Spatial-FST** only utilizes Equation (8) to aggregate spatial information and original local information by a residual connection (i.e., Figure (3a)). **Temporal-FST** only adopts Equation (12) to aggregate temporal information and spatial information by a residual connection (i.e., Figure (3b)). **Spatio-Temporal-FST** only uses Equation (8) and Equation (12) to aggregate temporal information with spatial information and spatial information with original local information by a residual connection, respectively (i.e., Figure (3c)). We use **Non-FST** as a base variant, and the module indicates that no residuals are added between the transformer blocks. We compare the above five variants of FST on a standard benchmark in Section 4.2 and observe that the FST achieves the best performance. Moreover, we find that FST introduces minor modifications to the residual connection but grants maximum benefits.

### 3.3 Text Embedding Learning

For learning the contextual relations between the words in the video description sentence \( s_i \), we adopt a BERT language representation model to encode the word sequence, which applies the bidirectional training of the transformer [49] to language modeling. It includes 12 layers of transformer blocks. Each block has 12 attention heads, and the hidden size is 768. Here, we take the hidden state of the per-token outputs of the last 2 layers to represent the information of the entire input sentence \( F_s \in \mathbb{R}^{d_t} \). Finally, we transform each sentence representation \( F_s \in \mathbb{R}^{d_t} \) into a
text embedding feature \( F_t \in \mathbb{R}^d \) by using a pointwise linear layer and an attention-aware feature aggregation layer \[17\].

### 3.4 Joint Embedding Learning

The purpose of joint embedding learning between video and textual features is to perform similarity comparisons. For a given video \( V_i \), the proposed framework extracts two types of embedding features – video embeddings \( F_v \) and relation embeddings \( F_r \). We calculate the similarity between videos and sentences in both embedding spaces. Specifically, for a given sentence \( T_i \), the similarity score with \( V_i \) is obtained by summing the cosine similarities between its text embedding features \( F_t \) and two such types of video embedding features. It is formulated as

\[
S(V_i, T_i) = \lambda \cdot \cosine(F_r, F_t) + (1 - \lambda) \cdot \cosine(F_v, F_t),
\]

where \( 0 \leq \lambda \leq 1 \) is a hyperparameter to balance the importance of two similarity scores. Based on the defined similarity score, we use a hinge-based triplet ranking loss to encourage the similarity score of matched videos and sentences to be larger than those of mismatched ones by considering matching text-video pairs as positives and by considering all other pairwise text-video combinations in the batch as negatives, formulated by

\[
L_r = \left[ \delta - S(V_i, T_i) + S(V_i, T_j) \right]_+ + \left[ \delta - S(V_i, T_i) + S(V_j, T_i) \right]_+,
\]

where \( 0 < \delta \leq 1 \) is the margin, the operator \([x]_+ = \max(x, 0)\), and \( S(\cdot, \cdot) \) is the similarity function. \((V_i, T_i)\) represents the positive pair, while \((V_i, T_j)\) and \((V_j, T_i)\) represent the negative pairs available in the mini-batch.

### 4 EXPERIMENTS

#### 4.1 Experimental Setup

**4.1.1 Dataset.** We evaluated the proposed BiC-Net model on three benchmarks: MSR-VTT, MSVD, and YouCook2. The MSR-VTT dataset \[61\] is the most widely used dataset for text-video retrieval. It contains 10,000 YouTube video clips with 20 different text captions. Following the settings in \[35, 61, 64\], we adopt three kinds of evaluation settings. For the 1k-A test set \[64\], we use 7k train+val videos \[36\] and 9k train+val videos for training \[16\] and report results. The MSVD dataset \[7\] contains 1,970 video clips from YouTube. Each video clip has around 40 descriptions in multiple languages. We only adopt English annotations in experiments. Following prior work \[50\], we separate the dataset into 1,200 clips for training, 100 clips for validation, and 670 clips for testing. The YouCook2 dataset \[72\] contains 2,000 cooking videos with 14,000 video clips. It covers 89 types of recipes. Each video clip is described by a textual sentence. Referring to \[36\], we evaluate the text-video clip retrieval task on the validation clips.

**4.1.2 Evaluation Metrics.** We employ the widely used median retrieval rank (MedR) and recall rate at top \( K \) (R@\( K \)) for assessing retrieval accuracy. MedR measures the median rank position among where true positives are returned. R@\( K \) measures the fraction of true positives being ranked at the top \( K \) returned results. Therefore, lower MedR scores indicate higher performance; in contrast, higher R@\( K \) scores indicate better performance.

**4.1.3 Implementation Details.** We sample 26 video frames from it with the same temporal duration between every two frames. In our experiments, the ILSVRC-2012-CLS \[45\] pre-trained InceptionResNetV2 \[48\] is adopted to extract 1536-D 2D features and the Kinetics \[24\] pre-trained I3D \[6\] to extract 1024-D 3D features. The number \( N \) of regions within a frame is 36, identical
to [1]. The dimension $d$ of region features extracted from ResNet-101 is 2048-D. The dimensionality of video-embedding vectors $F_v$ and relation-embedding vectors $F_r$ are set as 1024-D. For each sentence, we use pre-trained BERT to extract 1536-D word embedding and apply a pointwise linear layer and an attention-aware feature aggregation layer [17] to obtain 1024-D text-embedding vectors.

We implement our proposed model using PyTorch¹ and train it on 4 Tesla V100 GPUs. We train for 60 epochs using an Adam optimizer [25] with a mini-batch size of 64. On the MSR-VTT, MSVD, and YouCook2, the learning rates are set to 0.0002, 0.0004, and 0.0004, respectively. As for the layer number $L$ of the transformer block, we set it to 4, 2, and 4 on the MSR-VTT, MSVD, and YouCook2 datasets, respectively. In addition, the trade-off parameter $\lambda$ in Equation (14) and the margin $\delta$ in Equation (15) are set to 0.5 and 0.2, respectively.

4.1.4 Baselines. To justify the effectiveness of our BiC-Net, we compare it with the following 16 methods:

- **STG [46].** This is a fine-grained model, which constructs a spatio-temporal graph to enrich the video representations for joint space learning and learn fine-grained semantic contexts for text-video retrieval.
- **HGR [9].** This method is based on fine-grained cross-modal alignment, which models the relation between words by a graph convolutional network and projects text and videos into multi-grained latent spaces.
- **Mithun et al. [37].** This model is designed for video retrieval tasks, which takes multi-modal features as input by a fusion strategy and utilizes a weighted triplet ranking loss to better learn the video feature.
- **Miech et al. [36].** This classic video retrieval method introduces a large-scale dataset of 136 million video-text pairs and projects videos and text into a latent embedding space by a gating module.
- **DualEncoding [13].** This is a dual encoding framework, which designs a dual-encoding network to encode text and videos into multi-level embedding in a coarse-to-fine fashion.
- **CE [31].** This is a multi-modal fusion-based method, which utilizes a collaborative gating module to fuse rich multi-modal features from different pre-trained experts for better video representation.
- **MMT [16].** This is also a multi-modal fusion approach, which uses BERT for text representation and proposes a multi-modal transformer to jointly encode diverse modalities in videos for video representation.
- **TCE [62].** This is a text-augmented retrieval approach, which uses a latent semantic tree augmented encoder to represent text and a GRU with a multi-head self-attention mechanism to encode videos.
- **COOT [17].** This is a cross-model alignment approach, which proposes hierarchical modeling of videos and paragraphs.
- **ViSERN [15].** This is a fine-grained feature-based approach, which exploits reasoning between frame regions for measuring video-text similarity.
- **T2VLAD [56].** This is also a cross-model alignment approach that introduces a paradigm of global-local alignment based on NetVLAD [2] to perform video retrieval.
- **SUPPORT-SET [38].** This is a contrastive learning approach, which designs a generative objective to improve the instance discrimination limitations of contrastive learning.

¹http://www.pytorch.org
Table 1. Performance of Introducing Visual Relations (VRs) for Cross-modal Retrieval

| Method       | Text-to-Video | Video-to-Text |
|--------------|---------------|---------------|
|              | R@1 | R@5 | R@10 | MedR | R@1 | R@5 | R@10 | MedR |
| VG           | 32.9 | 65.8 | 79.7 | 3    | 32.1 | 65.2 | 77.6 | 3    |
| VR_{st}      | 29.8 | 64.4 | 77.2 | 3    | 29.2 | 63.6 | 76.8 | 3    |
| VG + VR_{m}  | 32.7 | 66.0 | 79.4 | 3    | 32.9 | 67.0 | 79.6 | 3    |
| VG + VR_{s}  | 33.8 | 69.3 | 82.9 | 2    | 36.2 | 72.4 | 84.3 | 2    |
| VG + VR_{t}  | 33.7 | 67.5 | 81.7 | 3    | 34.0 | 70.2 | 82.7 | 2    |
| VG + VR_{st} (BiC-Net) | **39.4** | **75.5** | **86.7** | **2** | **39.4** | **76.5** | **85.9** | **2** |

The evaluations are done on 1k-A test set for MSVR [16].

- **HGLMM FV** [36]. This is a classic approach that employs Fisher Vector [39] with Hybrid Gaussian Laplacian Mixture Model (HGLMM) distributions and CCA algorithm [23] for video-text retrieval.
- **AME** [27]. This is a fine-grained feature-based approach, which constructs a spatial and temporal graph for videos and designs an adversarial learning strategy to further narrow the domain gap between text and video representations.
- **Frozen** [4]. This is an end-to-end pre-training approach, which uses the recent ViT [2] as the visual encoder and designs a curriculum learning schedule to train the model on both image and video datasets.
- **CLIP4Clip-meanP** [33]. This is also an end-to-end retrieval approach, which uses CLIP [42] to extract the frame features and the text features, and then uses the mean pooling to aggregate the feature of all frames for video representation.

### 4.2 Ablation Studies

(1) **Experiments with Spatio-temporal Relations.** We experimented with variants of our model to verify the effectiveness of introducing spatio-temporal relations for text-video retrieval:

- **VG.** We only utilize the pre-trained 2D and 3D CNNs to extract the global features of the whole video as video embedding learning.
- **VG + VR_{m}**. We apply the average-pooling features of all regions without using the features extracted by fine-grained spatio-temporal transformer as relation embedding learning.
- **VG + VR_{s}**. We only utilize regional spatial relation features as relation embedding learning and global features as video embedding learning.
- **VG + VR_{t}**. We only utilize regional temporal relation features as relation embedding learning and global features as video embedding learning.
- **VR_{st}**. We only utilize regional spatio-temporal relation features as relation embedding learning.

We explore these model variants on the MSR-VTT, as shown in Table 1. We omit the results on MSVD and YouCook2 because of space limitations, but they show similar trends to MSR-VTT. From the results, we have the following observations. First, as expected, on both text-to-video and video-to-text, our BiC-Net, VG + VR_{s}, and VG + VR_{t} significantly outperform VG alone. The result verifies the significance of introducing spatio-temporal relation representation. Second, compared with VG + VR_{m}, the performance of our BiC-Net verifies that the fine-grained spatio-temporal transformer can capture the fine-grained local relational features. Third, compared with two variants of models (i.e., VG and VR_{st}) that only use either global visual features or local
Table 2. Performance of the Variants for FST for Cross-modal Retrieval

| Method                | Text-to-Video | Video-to-Text |
|-----------------------|---------------|---------------|
|                       | R@1 | R@5 | R@10 | MedR | R@1 | R@5 | R@10 | MedR |
| dataset split from [16]|      |      |       |      |      |      |       |      |
| Non-FST               | 36.2 | 73.9 | 84.4 | 2    | 38.2 | 74.3 | 87.5 | 2    |
| Spatial-FST           | 37.8 | 71.7 | 85.0 | 2    | 39.2 | 74.5 | 85.2 | 2    |
| Temporal-FST          | 37.8 | 71.9 | 85.2 | 2    | 40.1 | 73.9 | 85.8 | 2    |
| Spatio-Temporal-FST   | 38.2 | 73.2 | 85.8 | 2    | 39.3 | 74.2 | 85.5 | 2    |
| FST (BiC-Net)         | **39.4** | **75.5** | **86.7** | 2    | **39.4** | **76.5** | **85.9** | 2    |

1The evaluations are done on 1k-A test set (Training-9k) [16] for MSR-VTT.

Relational features, our model considers both global and local relational features to achieve the best performance. This verifies the effectiveness of aligning the global visual and local relational features with text features on two embedding spaces. Notably, global visual features and local relational features are highly complementary, and their combination leads to an improvement far beyond the performance of the global visual features alone. Moreover, compared with VG + VRs and VG + VRt, our BiC-Net achieves substantially better performance, which reveals the complementary nature of the spatial and temporal relation features.

(2) Evaluation of FST. We test the effectiveness of our proposed FST and its variants on relation embedding learning. As shown in Table 2, our FST and its variants achieve better performance than Non-FST, which indicates the effectiveness of spatio-temporal relation modeling by adding residual blocks. The difference between Spatial-FST and Temporal-FST is that a residual block is added at different positions. We can see that Temporal-FST significantly surpasses Spatial-FST, which indicates the importance of temporal relation modeling. Spatio-Temporal-FST adds a residual block based on Spatial-FST/Temporal-FST, which achieves better performance than Spatial-FST/Temporal-FST by aggregating temporal information with spatial information. In the end, compared with other variants, we observed that our proposed FST achieves the best performance when adding three residual blocks. This indicates that adding residual blocks simultaneously in our model can establish identity mapping between encoder blocks and enable gradients to flow directly across different encoder blocks to improve the model performance further. To sum up, the contribution of each component enables our FST to learn high-level spatio-temporal relation information.

4.3 Comparison with State-of-the-Art Methods

To demonstrate the effectiveness of the BiC-Net solution, we compared it to several state-of-the-art baselines: (1) RNN-based methods: DualEncoding [13] and TCE [62]; (2) multi-modal fusion methods: Mithun et al. [37], CE [31], and MMT [16]; (3) GCN-based methods: ViSERN, [15], STG [46], and AME [20]; (4) transformer-based methods: COOT [17], CLIP4Clip [33], and Frozen [4]; and (5) other methods: HGLMM FV CCA [36], Miech et al. [36], SUPPORT-SET [38], and T2VLAD [56].

(1) Experiments on MSR-VTT: The experimental results are presented in Table 3. We can observe that for all data partitions, our proposed method consistently outperforms all compared to traditional RNN-based methods and multimodal fusion methods in all evaluation metrics by a large margin, including CE [31], MMT [16], and T2VLAD [56], which use expert features (e.g., object, motion, face, scene, sound, and speech). Moreover, our BiC-Net significantly outperforms the recent spatio-temporal relation-based method (STG [46]) in all evaluation metrics. In particular, it boosts the text-video retrieval quality by a margin of 28.6% in R@10 on the full test set. This condition reveals the effectiveness of our BiC for modeling video global and relational information. The obvious performances are shown on the full test set and 1k-A test set (training-7k).
Table 3. Cross-modal Retrieval Comparison with State-of-the-Art Methods on MSR-VTT

| Method                  | Text-to-Video | Video-to-Text |       |       |       |
|-------------------------|---------------|---------------|-------|-------|-------|
|                         | R@1          | R@5          | R@10  | MedR  |       |
| Full test set [61]      |               |               |       |       |       |
| STG [46]                | 8.3          | 23.7          | 33.9  | 28    | -     |
| HGR [9]                 | 9.2          | 26.2          | 36.5  | 24    | 15.0  |
| DualEncoding [13]       | 11.6         | 30.3          | 41.3  | 17    | 22.5  |
| T2VLAD [56]             | 12.7         | 34.8          | 47.1  | 12    | 20.7  |
| BiC-Net                 | 19.2         | 47.0          | 62.5  | 6     | 20.6  |
| 1k-B test set [35]      |               |               |       |       |       |
| CE [31]                 | 18.2         | 46.0          | 60.7  | 7     | 18.0  |
| DualEncoding [13]       | 23.0         | 50.6          | 62.5  | 5     | 25.1  |
| MMT [16]                | 20.3         | 49.1          | 63.9  | 6     | 21.1  |
| T2VLAD [56]             | 26.1         | 54.7          | 68.1  | 4     | 26.7  |
| BiC-Net                 | 34.0         | 71.1          | 84.1  | 3     | 37.9  |
| 1k-A test set (training-7k) [36] |         |               |       |       |       |
| Miech et al. [36]       | 12.1         | 35.0          | 48.0  | 12    | -     |
| STG [46]                | 15.5         | 39.2          | 50.4  | 10    | -     |
| TCE [62]                | 17.1         | 39.9          | 53.7  | 9     | -     |
| DualEncoding [13]       | 21.6         | 49.5          | 62.3  | 6     | 27.8  |
| BiC-Net                 | 32.8         | 68.2          | 82.4  | 3     | 36.8  |
| 1k-A test set (Training-9k) [16] |         |               |       |       |       |
| MMT [16]                | 24.6         | 54.0          | 67.1  | 4     | 24.4  |
| SUPPORT-SET [38]        | 27.4         | 56.3          | 67.7  | 3     | 26.6  |
| Frozen [4]              | 31.0         | 59.5          | 70.5  | 3     | -     |
| CLIP4Clip-meanP [33]    | 44.5         | 71.4          | 81.6  | 2     | -     |
| BiC-Net                 | 39.4         | 75.5          | 86.7  | 2     | 39.4  |

We also compare with some typical transformer-based methods, such as CLIP4Clip [33] and Frozen [4]. CLIP4Clip adopts the language-vision transformer model of CLIP [42] pre-trained on a large-scale text-image dataset as a backbone. Frozen uses a transformer-based video model [5] as a backbone. In contrast, we design a new transformer-based backbone to model spatio-temporal relations and global temporal information. Our BiC outperforms most of the compared methods on 1k-A test set (Training-9k), e.g., BiC 86.7% vs. CLIP4Clip 81.6% with regard to text-to-video R@10. This indicates that learning cross-modal complementarity in a cooperative and complementary manner takes effect.

(2) Experiments on MSVD: Table 4 summarizes the performance comparison results. We also observe that our proposed BiC-Net outperforms recent state-of-the-art methods in terms
Table 5. Cross-modal Retrieval Comparison with State-of-the-art Methods on YouCook2

| Method                  | Text-to-Video | Video-to-Text |
|-------------------------|---------------|---------------|
|                         | R@1 | R@5 | R@10 | MedR | R@1 | R@5 | R@10 | MedR |
| HGLMM FV CCA [36]       | 4.6  | 14.3 | 21.6 | 75   | -   | -   | -    | -    |
| Miech et al. [36]       | 4.2  | 13.7 | 21.5 | 65   | -   | -   | -    | -    |
| COOT [17]               | 5.9  | 16.7 | 24.8 | 49.7 | -   | -   | -    | -    |
| AME [20]                | 7.6  | 21.5 | 32.8 | 28   | 7.9 | 22.5 | 32.2 | 28   |
| BiC-Net                 | 8.7  | 23.9 | 33.5 | 31   | 8.3 | 23.6 | 32.6 | 31   |

TS: trained from scratch on YouCook2.

of most indicators. Note that among all these methods, ViSERN [15] uses only local video features to compute the similarity between the video and text. Analogously, we also observe that BiC-Net outperforms the local feature-based method ViSERN [15] by a great margin. This reveals that jointly modeling the global and local video representation plays a significant role in text-video retrieval, contributing to more powerful representation. To ensure a fair comparison, we compare the previous state-of-the-art method, SUPPORT-SET [38], without pre-training on HowTo100M [36]. Under the full fair comparison, our BiC outperforms the previous best method SUPPORT-SET by 9.3% on video-text retrieval R@10. Notably, on MSVD, the performance of our model is not as outstanding. The reason for small gains is that the transformer has the property of lacking structural bias, making it prone to overfitting for small-scale data.

(3) Experiments on YouCook2: As shown in Table 5, our method achieves the best performance, which is 8.7% absolute gains in the evaluation metric of text-video retrieval R@10 better than COOT [17]. In Miech et al. [36] and COOT [17], the global video features are used for video representation, whose performances are worse than most methods. AME [20] adopts global features and handcrafted graph-based relation features. AME achieves better performances than Miech et al. [36], while their performances are worse than ours, which indicates that the global and local information learned by our method can be mutually promoted in a complementary manner. This observation indicates that in addition to the global video features, local relation features are also important for video representation.

4.4 Parametric Sensitivity Analysis

We carry out experiments to explore how the layer number of transformer block L and the trade-off parameter λ affect the retrieval performance. Note that we omit the video-text retrieval results on three datasets due to space limitations, which show similar trends to text-video retrieval. First, we analyze the influence of the hyperparameter, that is, the layer number of transformer blocks on the MSR-VTT 1k-A test set [16], MSVD, and YouCook2 datasets. Figure 4 (left) presents the results across the layer number of transformer blocks on the two datasets by R@10; note that R@1 and R@5 present the same trend, in which performances increase up to certain numbers (4, 2, and 4 for MSR-VTT, MSVD, and YouCook2 datasets, respectively) and then become stable. This result is due to the model’s capability of capturing the spatio-temporal relations of the deepest layer numbers.

The influence of the hyperparameter λ in Equation (11) is revealed in Figure 4 (right). We assign different trade-off parameters λ to the two scores (i.e., VRst and VG) to observe their influence on the matching performance of the three datasets. By analyzing the results shown in Figure 4 (right), we have the following observations: (1) The leftmost part of Figure 4 (right) shows the results when VRst accounts for 0, that is, when the proportion of VG is 1, which means that we remove the visual relations module from our model (i.e., VG). We can observe that when the spatio-temporal relations module is removed, the retrieval performance is reduced by a large margin over
Fig. 4. The figure on the left displays the R@10 results over various layer numbers of transformer blocks. The figure on the right shows the R@10 results over various different weight combinations of the global and relation similarities.

Table 6. Comparison of Different Models in Terms of Model Size and Computation Overhead at the Inference Stage

| Model           | Parameters (M) | FLOPs (G) |
|-----------------|----------------|-----------|
| MMT [16]        | 133.4          | 12.64     |
| DualEncoding [13]| 95.9           | 3.64      |
| BiC-Net         | 31.48          | 10.33     |

the three datasets. This condition shows the positive effect of comprehensively introducing spatio-temporal relations for text-video retrieval. In addition, it is worth noting that VG leads to a larger performance improvement for text-to-video retrieval R@10. This may be due to the fact that the retrieval performance benefits more from the global features in the target modality. (2) Increasing the VR_{st} proportion substantially boosts the model’s performance. Our model performs best on the three datasets when $\lambda = 0.5$. Therefore, we argue that VR_{st} and VG occupy the same contribution to the overall similarity. We conclude that the two similarities work together to obtain the best retrieval performance in a cooperative manner.

4.5 Model Complexity

We compare our method with open-source methods in terms of model size and computation overhead at the inference stage. As shown in Figure 4 (right), since the performance of using one layer of transformer block outperforms MMT by a large margin, we only calculate the model size and computational overhead of using one layer of transformer block. Analogously, we also observe that our BiC-Net with one layer of transformer block outperforms DualEncoding by a great margin on the MSR-VTT 1k-A test set [36]. Note that we omit the text-video retrieval results of VSR with a layer of transformer block on the MSR-VTT 1k-A test set [36] due to space limitations, which show similar trends to the MSR-VTT 1k-A test set [16]. In addition, we conclude that for each additional layer of transformer block, the computational cost will increase by 8.29 GFLOPs, and the parameters will increase by 25.19M. Following [13], we measure the number of FLOPs required for a text-video pair. As shown in Table 6 and Figure 4 (left), we have two main observations: (1) our
BiC-Net with one layer of transformer block achieves 85.6% text-to-video R@10 accuracy on the MSR-VTT 1k-A test set [16], which is 18.5% higher than MMT, with fewer parameters and lower computational cost. (2) our BiC-Net with one layer of transformer block is smaller and slightly slower than DualEncoding.

### 4.6 Qualitative Results

Figure 5 shows two examples of text-to-video retrieval results between the model with and without visual spatio-temporal relations. We specifically choose two text-video retrieval examples that include complex spatio-temporal relationships. Figure 5 (a) shows multiple sub-actions retrieval examples; the sentence describes two objects (“man” and “crab”) and two actions (“cutting open a crab” and “taking the meat out”) in a short-term segment, which requires accurate spatio-temporal grounding. Comparing BiC-Net with its variant VG, our model successfully retrieves the correct video, which contains all spatio-temporal relationships and entities described in the sentence. The second video only contains “taking the meat out from crab” actions. The third and fourth videos only involve a “cutting” action and similar objects (e.g., “man” and “knife”). The fifth video also only contains an action (“taking the meat”). In the left example, the VG model also retrieves similar scenes (e.g., similar man and cutting action) in the video. However, we observe that videos involving related elements are only ranked as the true positive in the top 3 positions. The performance of VG indicates that removing the fine-grained spatio-temporal relationships hurts the expressiveness of the video representation and further degrades the retrieval performance. Another example is shown in Figure 5(b), which requires fine-grained spatio-temporal relation grounding. The positive example contains a scenario involving two objects (“woman” and “makeup brush”) and a fine-grained action (“applies eye shadow”). Comparing these results, we observe that the variant VG retrieves a list of similar action videos, which cannot capture the fine-grained action (“applies eye shadow to her right eye”) in the video. Our model not only identifies the relevant objects “woman” and “makeup brush” but also captures the fine-grained relations between them. Again, it verifies the effectiveness of introducing spatio-temporal relation features to distinguish videos with the same visual components but with different relations. Our relation-augmented model still grapples with scenarios in which the text query encompasses both scenic descriptions and depictions of multiple objects. For instance, in Figure 6(a) on the left, despite the fact that the text query aims to retrieve the image with “mountains”, our model persistently ranks images depicting “a person and mountains” at the top. Similarly, in Figure 6(a) on the right, images containing mountains are retrieved. In Figure 6(b), where the text query aims
Fig. 6. Qualitative examples of the text-to-video tasks: In (a), (b), we show retrieval ranks of BiC-Net and the variant VG on MSR-VTT dataset test set [36]. Given a textual description as a query, we retrieve the most relevant video ranked from top to bottom. True positives are bounded in green boxes.

Fig. 7. Visualization of attention map on sample clips from the MSR-VTT. The top row presents original frames, and the bottom presents corresponding attention maps.

to retrieve images of “several young girls”, both our model and the baseline model continue to show a tendency to prioritize images showcasing “a group of people”.

4.7 Visualization Results
To intuitively observe the effectiveness of introducing spatio-temporal relations, we visualize the attention map between sentences and visual regions to infer the value of the spatio-temporal relation features. We select 2 videos, including two positive examples in text-to-video retrieval from MSR-VTT. In Figure 7, we show the original frames and attention maps. As can be seen, our BiC-Net learns to value core parts with intense semantic relations, such as “man + crab” in “cutting open a crab and taking the meat out”, and “woman + makeup brush” in “applies eye shadow to her right eye”. Furthermore, we find that the salient regions (e.g., man, crab, woman’s right eye) are highlighted separately in Figure 7. This also verifies that our model can learn fine-grained relational information with the corresponding text sentences.

5 CONCLUSIONS
This work contributes to a novel modeling method for cross-modal text-video retrieval. We claim that video representation should learn not only from global features but also from local fine-grained spatio-temporal relationships. To fulfill this target, we design the Bi-Branch Complementary Network (BiC-Net) to capture local relational and global visual information for modeling
comprehensively. Extensive experimental results on three benchmarks have demonstrated the effectiveness and superiority of our proposed method. However, we still face an inherent computational burden of attention in processing long-length videos with more complex local relations. Therefore, we leave computational optimization of the multi-layer spatio-temporal transformer as future work.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 6077–6086.

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5297–5307.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. arXiv:1607.06450 (2016).

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 1728–1738.

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space–time attention all you need for video understanding. In Proceedings of the IEEE Conference on Machine Learning. 813–824.

[6] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 6299–6308.

[7] David Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 190–200.

[8] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In Proceedings of the 24th ACM International Conference on Multimedia. 32–41.

[9] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10635–10644.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 4171–4186.

[11] Jianfeng Dong, Xirong Li, and Cees G. M. Snoek. 2018. Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia 20, 12 (2018), 3377–3388.

[12] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual encoding for zero-example video retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9346–9355.

[13] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 8 (2021), 4065–4080.

[14] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. arXiv:2106.11097 (2021).

[15] Zerun Feng, Zhimin Zeng, Caili Guo, and Zheng Li. 2020. Exploiting visual semantic reasoning for video-text retrieval. In International Joint Conference on Artificial Intelligence. 1005–1011.

[16] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In Proceedings of the European Conference on Computer Vision. 214–229.

[17] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. COOT: Cooperative hierarchical transformer for video-text representation learning. Advances in Neural Information Processing Systems 33 (2020), 22605–22618.

[18] Yutian Guo, Jingjing Chen, Hao Zhang, and Yu-Gang Jiang. 2020. Visual relations augmented cross-modal retrieval. In Proceedings of the 2020 Conference on Multimedia Retrieval. 9–15.

[19] Ning Han, Jingjing Chen, Guangyi Xiao, Hao Zhang, Yawen Zeng, and Hao Chen. 2021. Fine-grained cross-modal alignment network for text-video retrieval. In Proceedings of the ACM International Conference on Multimedia. 3826–3834.

[20] Ning Han, Jingjing Chen, Hao Zhang, Huanwen Wang, and Hao Chen. 2022. Adversarial multi-grained embedding network for cross-modal text-video retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications 18, 2 (2022), 1–23.

[21] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. 2022. Group contextualization for video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 928–938.

[22] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). arXiv:1606.08415 (2016).
[23] Harold Hotelling. 1936. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*. 321–377.

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv:1705.06950* (2017).

[25] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*. 1–15.

[26] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*. 1–14.

[27] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: ClipBERT for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7331–7341.

[28] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. 2020. SEA: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia* 23 (2020), 4351–4362.

[29] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. 2021. HiT: Hierarchical Transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11915–11925.

[30] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Zechao Li, Qi Tian, and Qingming Huang. 2022. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3003–3018.

[31] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. In *British Machine Vision Conference*. 279.

[32] Wei Lu, Desheng Li, Liqiang Nie, Peiguang Jing, and Yuting Su. 2021. Learning dual low-rank representation for multi-label micro-video classification. *IEEE Transactions on Multimedia* 25 (2021), 77–89.

[33] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.

[34] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9876–9886.

[35] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv:1804.02516* (2018).

[36] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2630–2640.

[37] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 19–27.

[38] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G. Hauptmann, Joao F. Henriques, and Andrea Vedaldi. 2021. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*. 1–18.

[39] Florent Perronnin and Christopher Dance. 2007. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[40] Mengshi Qi, Jie Qin, Yi Yang, Yunhong Wang, and Jiebo Luo. 2021. Semantics-aware spatial-temporal binaries for cross-modal video retrieval. *IEEE Transactions on Image Processing* 30 (2021), 2989–3004.

[41] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shiliang Pu, and Jun Xiao. 2019. Video relation detection with spatio-temporal graph. In *Proceedings of the ACM International Conference on Multimedia*. 84–93.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. 8748–8763.

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* 28 (2015).

[44] Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. 2021. AVLNet: Learning audio-visual-language representations from instructional videos. In *Annual Conference of the International Speech Communication Association*. 1584–1588.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.
[46] Xue Song, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. 2021. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia* 24 (2021), 2914–2923.

[47] Zixuan Su, Xindi Shang, Jingjing Chen, Yu-Gang Jiang, Zhiyong Qiu, and Tat-Seng Chua. 2020. Video relation detection via multiple hypothesis association. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3127–3135.

[48] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4278–4284.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances on Neural Information Processing Systems*. 6000–6010.

[50] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence — video to text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4534–4542.

[51] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. 2020. Dual path interaction network for video moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4116–4124.

[52] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Xuejin Chen, and Jiebo Luo. 2023. Semantic and relation modulation for audio-visual event localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 7711–7725.

[53] Hao Wang, Zheng-Jun Zha, Liang Li, Jiebo Luo, and Dong Liu. 2021. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7026–7035.

[54] Wei Wei, Junyu Gao, Xiaoshan Yang, and Changsheng Xu. 2022. Many hands make light work: Transferring knowledge from auxiliary tasks for video-text retrieval. *IEEE Transactions on Multimedia* 25 (2022), 2661–2674.

[55] Xiaolong Wang and Abhinav Gupta. 2018. Videos as space-time region graphs. In *European Conference on Computer Vision*. 399–417.

[56] Xiaohua Wang, Linchao Zhu, and Yi Yang. 2021. T2VLAD: Global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5079–5088.

[57] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 450–459.

[58] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. 2019. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9964–9974.

[59] Peng Wu, Xiangteng He, Mingqian Tang, Yiliang Lv, and Jing Liu. 2021. HANet: Hierarchical alignment networks for video-text retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3518–3527.

[60] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. 2020. Visual relation grounding in videos. In *Proceedings of the European Conference on Computer Vision*. 447–464.

[61] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5288–5296.

[62] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1339–1348.

[63] Xun Yang, Meng Wang, Richang Hong, Qi Tian, and Yong Rui. 2017. Enhancing person re-identification in a self-trained subspace. *ACM Transactions on Multimedia Computing, Communications, and Applications* 13, 3 (2017), 1–23.

[64] Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision*. 471–487.

[65] Yawen Zeng, Ning Han, Keyu Pan, and Qin Jin. 2023. Temporally language grounding with multi-modal multi-prompt tuning. *IEEE Transactions on Multimedia* (2023), 1–12.

[66] Yawen Zeng, Qin Jin, Tengfei Bao, and Wenfeng Li. 2023. Multi-modal knowledge hypergraph for diverse image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3376–3383.

[67] Yawen Zeng, Keyu Pan, and Ning Han. 2023. RewardT2LG: Learning to temporally language grounding from flexible reward. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2344–2348.

[68] Hao Zhang, Lechao Cheng, Yanbin Hao, and Chong-wah Ngo. 2022. Long-term leap attention, short-term periodic shift for video classification. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5773–5782.

[69] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. 2021. Token shift transformer for video classification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 917–925.

[70] Yanchao Zhang, Weiqing Min, Liqiang Nie, and Shuqiang Jiang. 2020. Hybrid-attention enhanced two-stream fusion network for video venue prediction. *IEEE Transactions on Multimedia* 23 (2020), 2917–2929.
[71] Rui Zhao, Kecheng Zheng, and Zheng-jun Zha. 2020. Stacked convolutional deep encoding network for video-text retrieval. In 2020 IEEE International Conference on Multimedia and Expo. 1–6.

[72] Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In Proceedings of the AAAI Conference on Artificial Intelligence. 7590–7598.

[73] Xiaofeng Zou, Kenli Li, and Cen Chen. 2020. Multilevel attention based U-shape graph neural network for point clouds learning. IEEE Transactions on Industrial Informatics 18, 1 (2020), 448–456.

Received 11 September 2022; revised 19 August 2023; accepted 24 September 2023