Multi-Feature Graph Attention Network for Cross-Modal Video-Text Retrieval

Xiaoshuai Hao
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
haoxiaoshuai@iie.ac.cn

Yucan Zhou*
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
zhouyucan@iie.ac.cn

Dayan Wu
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
wudayan@iie.ac.cn

Wanqian Zhang
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
zhangwanqian@iie.ac.cn

Bo Li
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
libo@iie.ac.cn

Weiping Wang
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
wangweiping@iie.ac.cn

ABSTRACT

Cross-modal retrieval between videos and texts has attracted growing attention due to the rapid growth of user-generated videos on the web. To solve this problem, most approaches try to learn a joint embedding space to measure the cross-modal similarities, while paying little attention to the representation of each modality. Video is more complicated than the commonly used visual feature, since the audio and caption on the screen also contain rich information. Recently, the aggregations of multiple features in videos boost the benchmark of the video-text retrieval system. However, they usually handle each feature independently, which ignores the interchange of high-level semantic relations among these multiple features. Moreover, despite the inter-modal ranking constraint where semantically-similar texts and videos should stay closer, the modality-specific requirement, i.e. two similar videos/texts should have similar representations, is also significant. In this paper, we propose a novel Multi-Feature Graph ATtention Network (MFGATN) for cross-modal video-text retrieval. Specifically, we introduce a multi-feature graph attention module, which enriches the representation of each feature in videos with the interchange of high-level semantic information among them. Moreover, we elaborately design a novel Dual Constraint Ranking Loss (DCRL), which simultaneously considers the inter-modal ranking constraint and the intra-modal structure constraint to preserve both the cross-modal semantic similarity and the modality-specific consistency in the embedding space. Experiments on two datasets, i.e. MSR-VTT and MSVD, demonstrate that our method achieves significant performance gain compared with the state-of-the-arts.

CCS CONCEPTS

• **Information systems → Video search**.

KEYWORDS

Video-Text Retrieval; Multi-Feature Aggregation; Graph Attention Network

ACM Reference Format:
Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, and Weiping Wang. 2021. Multi-Feature Graph Attention Network for Cross-Modal Video-Text Retrieval. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3460426.3463608

1 INTRODUCTION

With the exponential growth of user-generated videos on the Internet, cross-modal retrieval between video data and natural language descriptions, known as video-text retrieval, has attracted much attention. The goal of video-text retrieval is to retrieve and rank the videos in the database according to the query text given by users. To achieve it, the current dominant paradigm for video-text retrieval [9, 10, 26, 30] tries to map the queries and the videos into a joint embedding space, where the semantically-similar texts and videos are much closer to each other and vice versa.

Most existing methods are adopted from the image-text embedding methods, which focus on the visual representation of videos. Some researchers [4, 5, 7, 16, 31, 32, 32, 35, 40, 42, 43] struggle to find a representative video frame, and then feed it into the image-text model for video-text retrieval. However, other rich information in the videos effective for video-text retrieval is ignored. Given a query like 'a little girl reacting to a video of President Obama giving a speech', satisfactory results are difficult to be retrieved without the audio or the caption on the screen.

*Corresponding author.

ICMR '21, August 21–24, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8463-6/21/08…$15.00
https://doi.org/10.1145/3460426.3463608

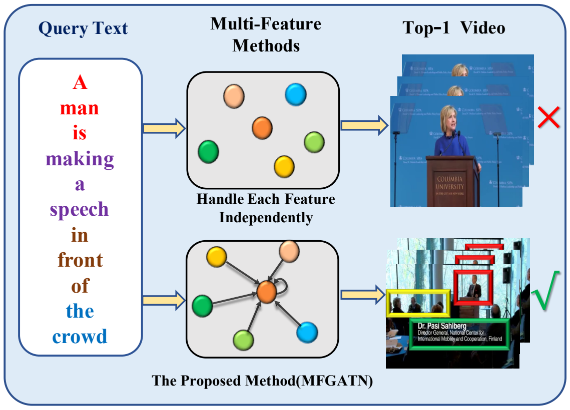


Figure 1: Illustration of the differences between handle each feature independently and the proposed method. The proposed method considers the interchange of high-level semantic information among multiple features and successfully retrieves the correct video given an complex query.

Recently, feature aggregation methods greatly boost the benchmark of video-text retrieval, which make use of different features in videos like object, motion, audio, and caption on the screen. However, they usually handle each feature independently, which ignores the interchange of high-level semantic information among these multiple features. Leveraging the interchange information among different features is of great importance to build effective video representations. As illustrated in Figure 1, given a complex query like 'A man is making a speech in front of the crowd', neither of the features 'appearance', 'motion' or 'audio' can fully describe the scene. On the contrary, when these features are processed together, the higher-level semantics can be obtained. How to fully exploit the rich and heterogeneous information in videos is still an open problem, which is also the primary motivation of this paper.

Moreover, all existing works train the embedding network by considering the inter-modal constraint to make the semantically-similar texts and videos much closer to each other and vice versa. Ideally, a good embedding space should also satisfy the requirement that similar videos/texts should stay closer. Thus, we argue that preserving this modality-specific characteristic is essential for learning the embedding space.

In this paper, we propose a novel Multi-Feature Graph ATTention Network (MFGATN). Specifically, we devise a multi-feature graph attention module, which enriches the representation of each feature with the interchange of high-level semantic information among multiple features. Besides, we elaborately design a novel Dual Constraint Ranking Loss (DCRL) that simultaneously considers the inter-modal ranking constraint and the intra-modal structure constraint. In light of the proposed DCRL, we can preserve the modality-specific characteristics in the embedding space to further improve retrieval performance. With our MFGATN, not only more target videos can be retrieved, but also similar videos are ranked higher than other irrelevant videos as they are mapped closer in the embedding space. To show the effectiveness of the proposed MFGATN, we conduct experiments on two benchmark datasets. The MFGATN method achieves 21% and 17.6% relative improvements on R@1 compared with the state-of-the-art method on the MSR-VTT and the MSVD datasets, respectively. The main contributions of this work can be summarized as follows:

- We propose a novel Multi-Feature Graph Attention Network to aggregate multiple features in videos. By interchanging information among them, we can obtain more effective video representations.
- We elaborately design a novel Dual Constraint Ranking Loss (DCRL) that simultaneously considers the inter-modal ranking constraint and the intra-modal structure constraint, which makes both the semantically-similar video-text and the similar samples in each modality stay closer in the embedding space. To our best knowledge, this is the first loss function in video-text retrieval to preserve modality-specific characteristics.
- Our method achieves 21% and 17.6% relative improvements on R@1 compared with the state-of-the-art method on the MSR-VTT and the MSVD datasets, respectively.

2 RELATED WORK

2.1 Image-Text Retrieval

Recently, there has been increasing interest in learning robust visual-text embeddings for image-text retrieval [8–10, 12, 14, 18, 19, 26, 29, 36, 41]. Frome et al. [9] firstly propose a method to project words and visual contents into a joint space by a ranking loss that punishes the condition when a unmatched word is ranked higher than the matched one. Faghri et al. [8] modify the pairwise ranking loss based on violations caused by the hard-negatives (i.e., unmatched query closest to each training query) and has been shown to be effective in the retrieval task. Kiros et al. [14] extend the framework to encode images with CNN and sentences with RNN. Then, the following image-text retrieval methods adopt a similar approach with slight modifications in the input representations. In [26], authors propose a multi-modal attention mechanism to attend to sentence fragments and image regions selectively for similarity calculation. To enrich global representations, Gu et al. [10] further incorporate image and caption generation in a multi-task framework.

2.2 Video-Text Retrieval

Concept-based approaches [33, 38, 39] extract relevant concepts from queries and videos, and accordingly establish associations between these two modalities. The majority of the top-ranked solutions for TRECVID challenge belong to concept-based approaches [15, 21]. However, it is usually ineffective for complex long queries, since it is very difficult to describe the rich sequential information within both videos and queries using a few selected concepts.

Embedding-based approaches [1, 7, 11, 13, 17, 20, 25, 28] try to directly encode videos and texts into a common space. Many of these existing approaches [4, 7, 17] are inspired by the image-text embedding methods. Dong et al. [7] propose a dual multi-level encoding for both videos and queries. Chen et al. [4] propose a Hierarchical Graph Reasoning (HGR) model, which decomposes video-text into global-to-local levels. However, these methods do not take advantage of the rich and diverse information presented in

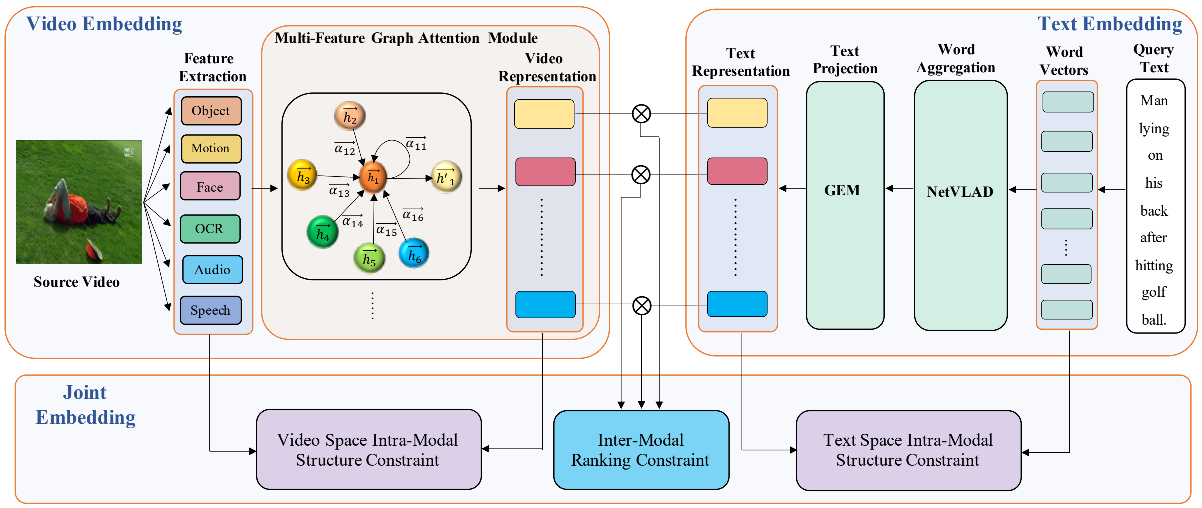


Figure 2: The Framework of Multi-Feature Graph Attention Network for video-text retrieval. Our proposed framework consists of three components: 1) the video embedding component adopts multi-feature graph attention module to enrich the representation of each feature with the interchange of high-level semantic information among multiple features. 2) the text embedding component encodes the query sentence into a single vector representation, and then projects it to separated subspaces for each video feature. 3) the joint embedding component learns the final multi-modal embedding space with the the inter-modal ranking constraint and the intra-modal structure constraint.

videos, such as objects, actions, faces, their combinations. Recently, it is widely studied that how to effectively aggregate the multiple features available in videos into a compact video representation. Mithun et al. [25] propose the JEMC framework using action, object, text and audio features to compute three corresponding text-video similarities. Miech et al. [23] propose a new model for learning a joint text-video embedding called Mixture-of-Embedding-Experts (MoEE), where the overall similarity is obtained as a weighted sum of each expert's similarity. Liu et al. [20] further adopt all video features and use a collaborative gating mechanism for modulating each expert feature according to the other experts. However, most of the existing methods ignore the interchange of high-level semantic information among multiple features, which is the major concern of our work.

2.3 Loss Function

Many prior methods require the inter-modal ranking constraint to make the semantically-similar texts and videos much closer to each other and vice versa. Miech et al. [23] adopt bi-directional max-margin ranking loss (Bi-MMRL) [20, 23] to train the video-text cross-modal embedding network. They minimize a hinge-based triplet ranking loss combined with the bi-directional ranking terms, which maximizes the similarity between a video embedding and the corresponding text embedding, and at the same time, minimizes the similarity to all other unmatched ones. Recently, focusing on hard-negatives is effective in many embedding tasks[8, 27]. Inspired by this, a few methods adopt bi-directional hard-negatives ranking loss (Bi-HNRL) [4, 7, 25] for this task to emphasize the hardest negatives, where the penalties incurred by the hardest negatives instead of all the negatives are considered. However, only considering the inter-modal ranking constraint will lead to a decrease in modality-specific characteristics. To address this problem, we elaborately design a novel dual constraint ranking loss function that simultaneously considers the inter-modal ranking constraint and the intra-modal structure constraint.

3 METHODOLOGY

Given a video V and a query text T, we try to create a pair of functions $\phi(V)$ and $\psi(T)$ mapping videos and texts into a joint embedding space, in which embeddings for matched texts and videos should lie close together, while embeddings for mismatched texts and videos should lie far apart. As illustrated in Figure 2, our proposed framework consists of three components: 1) the video embedding component extracts multiple features of videos and obtains fixed-length video feature vectors by temporal aggregation module, and then leverages multi-feature graph attention module to enrich the representation of each feature with the interchange of high-level semantic information among multiple features. 2) the text embedding component encodes the query sentence into a single vector representation, and then projects it to separated subspaces for each video feature. 3) the joint embedding component learns the final multi-modal embedding space with the the inter-modal ranking constraint and the intra-modal structure constraint.

3.1 Video Embedding

Feature Extraction: In order to make full use of the information in one video, we draw on a collection of pre-trained models to extract different video features. These operations map the video to a collection of M video feature embeddings $\left\{I_{var}^{(1)}, ..., I_{var}^{(M)}\right\}$. $I_{var}^{(i)}$ represents the i-th video feature (subscript var denotes a variable-length output when applied to a sequence of frames). In this paper, we set M=6, and extract features for object, motion, audio, speech, OCR , face. We use the features publicly released by [20]. Note that our method can be easily extended to more features if required. Each element of this collection is then aggregated along its temporal dimension, producing a fixed-length embedding per video $\left\{I^{(1)}, .., I^{(M)}\right\}$. For temporal aggregation function, we adopt a simple approach to aggregate the features. For object, motion, face embeddings, we average the frame-level features along the temporal dimension to produce a single feature vector per video. For speech, audio, OCR features, we adopt the NetVLAD mechanism proposed by Arandjelovic [2], which has been proven effective for the retrieval task [20].

Multi-Feature Graph Attention Module: Once the time aggregated embeddings are obtained, we apply linear projections to transform these embeddings into the same dimensionality. These projected video feature embeddings can be written as:

$$\boldsymbol{H} = \{h_1, h_2, ..., h_M\}, \tag{1}$$

where $h_i \in \mathbb{R}^F$, and F is the number of features.

To aggregate these multiple features, we first construct a multi-feature graph for each video. Specially, we assume that each video is represented by a set of nodes, $\boldsymbol{H} = \{h_1, ..., h_M\}$. Each node stands for a high-level video feature. To enrich the representation of each feature with the interchange of high-level semantic information among multiple features, we propose a multi-feature graph attention module (MFGAT).

Once a multi-feature graph is obtained, we then perform self-attention on the nodes−a shared attentional mechanism $\mathbf{a} : \mathbb{R}^F \times \mathbb{R}^F \rightarrow \mathbb{R}$ computes attention coefficients:

$$e_{ij} = \mathbf{a}\left(h_i, h_j\right), \tag{2}$$

which shows the significance of node j to node i. In this paper, the module allows every node to attend on all the nodes.

We perform masked attention on the graph structure—we only compute e_{ij} for all the neighbors of node i in the graph. To make the coefficients easily comparable across different nodes, we normalize them using the softmax function:

$$\alpha_{ij} = \text{softmax}_j\left(e_{ij}\right) = \frac{exp\left(e_{ij}\right)}{\sum_{k \in N_i} exp\left(e_{ik}\right)}. \tag{3}$$

where N_i are the neighbors of node i in the graph.

In our model, the proposed attention mechanism $\mathbf{a}$ is a single−layer neural network, which can be easily parametrized with a weight vector $\vec{a} \in \mathbb{R}^{2F}$, as well as the LeakyReLU non-linearity. The coefficients computed by the attention mechanism then is expressed as:

$$\alpha_{ij} = \frac{exp\left(LeakyReLU\left(\vec{a}^T\left[h_i \parallel h_j\right]\right)\right)}{\sum_{k \in N_i} exp\left(LeakyReLU\left(\vec{a}^T\left[h_i \parallel h_k\right]\right)\right)}, \tag{4}$$

where $\parallel$ denotes the concatenation operation.

Once the attention coefficients obtained, they are used to compute a linear combination of the features propagated to them, to produce the final output features for each node:

$$h_i' = \sigma\left(\sum_{j \in N_i} \alpha_{ij} h_j\right) + h_i. \tag{5}$$

Based on it, we obtain the new video features $V = \left\{h_i'\right\}_{i=1}^M$, which are enriched with the interchange of high-level semantic information among multiple features. The final video representation is then obtained by passing the modulated response of each video feature embedding through a Gated Embedding Module (GEM) [22] before concatenating the outputs together into a single fixed-length vector.

3.2 Text Embedding

Given a query sentence, we first propagate each word into the word2vec [24] model trained by Google News[1] to achieve their word embeddings. Then, all the word embeddings are passed through a pre-trained OpenAI-GPT model to extract the context-specific word embeddings. These word embeddings are then aggregated into a single sentence vector to obtain the entire sentence embedding using the NetVLAD [2] aggregation module. After the aggregation, we project the aggregated sentence vector to the separated subspaces for each video feature using Gated Embedding Module (GEM) [22]. The text representation then consists of M embeddings, represented by $T = \left\{\psi^i\right\}_{i=1}^M$.

3.3 Joint Embedding Learning

In this subsection, we introduce the Dual Constraint Ranking Loss (DCRL) in detail, which simultaneously considers the inter-modal ranking constraint and the intra-modal structure constraint.

Inter-modal ranking constraint: Existing works train the embedding network with the only consideration of the ranking constraints between modals, which makes the semantically similar texts and videos become closer and vice versa. While bridging the gap between an anchor and a positive sample, inter-modal ranking constraint can also maximize the distance between an anchor and a negative sample. The expression of the inter-modal ranking constraint of a video is as follows:

$$d\left(V_i, T_i\right) + m < d\left(V_i, T_j\right), \tag{6}$$

where, V_i (anchor) and T_i (positive sample) are the feature embeddings in the joint embedding space for the i-th video and text. T_j (negative sample) refers to the j-th text. $d\left(V, T\right)$ indicates the distance between two feature embeddings in the joint embedding space, and m indicates a margin constant. Analogously, given a text input, we set the inter-modal ranking constraint as follows:

$$d\left(T_i, V_i\right) + m < d\left(T_i, V_j\right). \tag{7}$$

In this triplet selection, there are two methods: the bi-directional max-margin ranking loss (Bi-MMRL), which calculates for all negatives; and the bi-directional hard-negatives ranking loss (Bi-HNRL) only the penalty incurred by the hardest negatives is considered. We adopt the Bi-HNRL as it has been proved more effective [7].

Intra-modal structure constraint: During the whole training procedure, if we only utilize the inter-modal ranking constraint, inherent characteristics within each modality (i.e., modality-specific characteristics) will be lost. To solve this problem, we devise a novel intra-modal structure constraint.

Suppose there are three samples (videos or texts), we can extract features using the process described in Section. 3.1 or Section. 3.2.

[1] https://code.google.com/archive/p/word2vec/

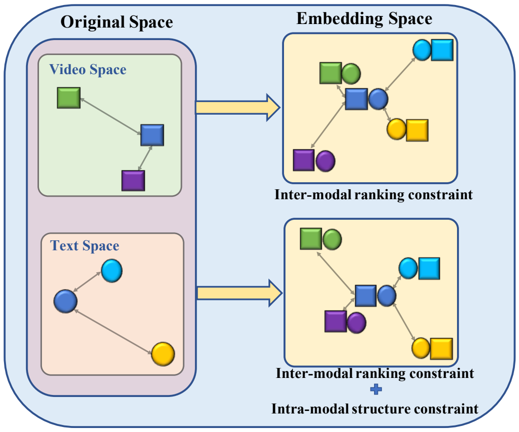


Figure 3: Embedding space with (top) and without (bottom) intra-modal structure constraint. By leveraging the intra-modal structure constraint, we can preserve modality-specific characteristics in the joint embedding space (best viewed in color).

Since deep features are not been fed into the joint embedding network, they can be used measure the modality-specific similarities. As Fig. 3 show, in the video space, blue is more similar to purple than green. In the text space, blue is more similar to sky-blue than yellow. By leveraging the intra-modal structure constraint in the embedding space, we can preserve modality-specific characteristics after the joint embedding process. The intra-modal structure constraint between samples is a soft relationship. When defining the intra-modal structure constrain, we do not use the margin constant. The expression of our proposed intra-modal structure constraint for a video is as follows:

$$d\left(V_i, V_j\right) < d\left(V_i, V_k\right), \;\; if\; d\left(\widetilde{V_i}, \widetilde{V_j}\right) < d\left(\widetilde{V_i}, \widetilde{V_k}\right), \tag{8}$$

where V_i, V_j, V_k are the video embeddings in the joint embedding space from i-th, j-th and k-th video, respectively. $\widetilde{V_i}$, $\widetilde{V_j}$, $\widetilde{V_k}$ are the video features from i-th, j-th and k-th in the original video space. Analogously, given a text input, we set the intra-modal structure constraint as follows:

$$d\left(T_i, T_j\right) < d\left(T_i, T_k\right), \;\; if\; d\left(\widetilde{T_i}, \widetilde{T_j}\right) < d\left(\widetilde{T_i}, \widetilde{T_k}\right), \tag{9}$$

where T_i, T_j, T_k are the text embeddings in the joint embedding space from i-th, j-th and k-th text, respectively. $\widetilde{T_i}$, $\widetilde{T_j}$, $\widetilde{T_k}$ are the text features from i-th, j-th and k-th text in the original text space.

Dual Constraint Ranking Loss (DCRL): Here, we can propose a simple yet effective ranking loss by the combination of the inter-modal ranking constraint and the proposed intra-modal structure constraint.

Assume there are one batch of text-video pairs, we have N pairs of embedded features (V_i, T_i). Here, V_i and T_i are the feature embeddings for the video and text in the i-th text-video pair in the joint embedding space. In light of the inter-modal ranking constraint, two difference types of triplets $\left(V_i, T_i, T_j\right)$ and $\left(T_i, V_i, V_j\right)$ can be constructed, where $i \neq j$. For the intra-modal structure constraint, we adopt two difference types of triplets $\left(V_i, V_j, V_k\right)$ and $\left(T_i, T_j, T_k\right)$, where $i \neq j \neq k$. Taking all these triplets into consideration, the Dual Constraint Ranking Loss (DCRL) can be written as:

$$\begin{aligned} L = & \sum_{i \neq j} max\left(0,\; V_i^T T_j - V_i^T T_i + m\right) \\ & + \sum_{i \neq j} max\left(0,\; T_i^T V_j - T_i^T V_i + m\right) \\ & + \lambda \left[\sum_{i \neq j \neq k} C_{ijk}\left(V\right)\left(V_i^T V_j - V_i^T V_k\right) \right. \\ & \left. + \sum_{i \neq j \neq k} C_{ijk}\left(T\right)\left(T_i^T T_j - T_i^T T_k\right) \right], \end{aligned} \tag{10}$$

where, λ balance the impact of intra-modal structure constraint. The function $C(\cdot)$ in Eq. 10 can be written as::

$$C_{ijk}\left(x\right) = sign\left(x_i^T x_k - x_i^T x_j\right) - sign\left(\widetilde{x}_i^T \widetilde{x}_k - \widetilde{x}_i^T \widetilde{x}_j\right), \tag{11}$$

where x_i, x_j and x_k are the feature embeddings in the joint embedding space and $\widetilde{x}_i$, $\widetilde{x}_j$ and $\widetilde{x}_k$ are intra-modal features in the original space. As stated above, the intra-modal structure constraint is soft. Hence, we replace real distance values with the *sign* function when introducing the intra-modal structure constrain Eq. 8 and Eq. 9 to the final loss funtion.

4 EXPERIMENTS

In this section, we first describe the datasets and evaluation metric in Sec. 4.1. Then, we describe the implementation details in Sec. 4.2. A comprehensive comparison of two video-text retrieval benchmark datasets is reported in Sec. 4.3. We report the ablation studies to further demonstrate the efficiency of our method in Sec. 4.4. Finally, extensive qualitative results are also presented in Sec. 4.5.

4.1 Datasets and Evaluation Metrics

MSR-VTT: The MSR-VTT [34] is a benchmark dataset for video-text retrieval. Originally developed for video captioning, the MSR-VTT dataset consists of $10k$ web video clips and $200k$ natural sentences which describe the semantic content of the clips. Each clip is assigned with 20 sentences. For a fair comparison, we follow the same data partitions in [25], which is the first work reporting video retrieval performance on the MSR-VTT dataset. Specifically, we use 6,513 clips for training, 497 clips for validation, and the remaining 2,990 clips for testing.

MSVD: The MSVD [3] dataset contains 1,970 Youtube clips, of which each video is annotated with about 40 sentences. For a fair comparison, we use the same data splits utilized in prior works [20], with 1,200 videos for training, 100 videos for validation, and 670 videos for testing.

Evaluation Metrics To evaluate the proposed method on the video-text retrieval task, we adopt the widely used evaluation metrics in most previous methods. R@K and Median Rank (Med R) are adopted to measure the rank-based performance. R@K denotes the percentage of test queries, which means one relevant item at least can be found among the top-K returned results. In this paper, we report results for R@1, R@5, and R@10. Med R denotes the median rank of the first relevant item in the returned results. Results with higher R@K and lower Med R are better. Moreover, the sum of R@1,

Table 1: Video-to-Text and Text-to-Video retrieval results on the MSR-VTT dataset. The proposed method performs the best.

Method	Text-to-Video Retrieval				Video-to-Text Retrieval				rsum
	R@1	R@5	R@10	Med R	R@1	R@5	R@10	Med R	
Single-feature method									
VSE [8]	5.0	16.4	24.6	47	7.7	20.3	31.2	28	105.2
VSE++ [8]	5.7	17.1	24.8	65	10.2	25.4	35.1	25	118.3
W2VV [6]	6.1	18.7	27.5	45	11.8	32.1	42.4	16	138.6
Dual Encoding [7]	7.7	22.0	31.8	32	13.0	30.8	43.3	15	148.6
HGR [4]	9.2	26.2	36.5	24	15.0	36.7	48.8	11	172.4
Multi-feature method									
JEMC [25]	7.0	20.9	29.7	38	12.5	28.9	39.1	21	138.1
Simple Concatenation	9.4	26.9	37.9	20	15.1	38.0	51.0	10	178.3
MoEE [23]	9.7	28.7	40.6	17	14.8	40.9	54.8	8	189.5
CE [20]	10.0	28.8	40.4	16	16.5	43.5	56.8	7.5	196.0
MFGATN	**12.1**	**32.9**	**45.2**	**13**	**21.4**	**51.2**	**64.8**	**5**	**227.6**

Table 2: Video-to-Text and Text-to-Video retrieval results on the MSVD dataset. The proposed method performs the best.

Method	Text-to-Video Retrieval				Video-to-Text Retrieval				rsum
	R@1	R@5	R@10	Med R	R@1	R@5	R@10	Med R	
Single-feature method									
ST [14]	2.6	11.6	19.3	51	2.99	10.9	17.5	77	64.89
LJRV [37]	7.7	23.4	35.0	21	9.85	27.1	38.4	19	141.45
VSE [8]	12.3	30.1	42.3	14	15.8	30.2	41.4	12	172.1
VSE++ [8]	15.4	39.6	53.0	9	21.2	43.4	52.2	9	224.8
Multi-feature method									
JEMC [25]	20.3	47.8	61.1	6	**31.5**	51.0	61.5	5	273.2
Simple Concatenation	18.2	45.1	60.4	7	20.6	48.2	59.0	6	251.5
MoEE [23]	19.1	46.9	62.4	6	23.1	51.9	62.8	5	266.2
CE [20]	19.3	47.2	62.6	6	23.4	50.4	61.5	5.5	264.4
MFGATN	**22.7**	**55.1**	**69.3**	**4**	27.8	**55.8**	**66.9**	**4**	**297.6**

R@5, and R@10, noted as rsum is also reported. The rsum is used to compare the overall performance.

4.2 Implementation Details

The MFGATN is implemented with the open resource framework PyTorch. We adopt the Adam optimizer for all our experiments and the margin of the inter-modal ranking loss is set to 0.3. We adopt Bi-HNRL as the inter-modal ranking loss and the hyper-parameter λ of the intra-modal structure loss is discussed detailly in section 4.4. Inspired by other baselines, we also freeze our pre-trained models for video feature extraction. All aggregated video features are projected to the same size as 768 before fed into the MFGAT module (i.e., F=768). For MSR-VTT, we train the model with a batch size of 64, a learning rate of 0.01, a weight decay of 5E-5. For MSVD, we set the batch size to 16, learning rate to 0.001, weight decay to 5E-5. After every epoch, we evaluate the proposed model on the validation set, and the final model is defined as the model with the best recalls.

4.3 Performance Comparisons

With the same settings and data partition, we compare the proposed MFGATN method with several state-of-the-art methods to demonstrate the efficacy. Video-text retrieval approaches can be divided into two categories according to the features for videos: single-feature methods and multi-feature methods. For single-feature methods, we compare with VSE [8], VSE++ [8], W2VV [6], dual encoding [7], HGR [4], LJRV [37], and ST [14]. Besides, we also compare it with several multi-feature methods, including JEMC [25], Simple Concatenation, MoEE [23],and CE [20]. The simple concatenation method connects multiple features to a single high-dimensional embedding, followed by a GEM.

For single-feature methods, we directly report the results from corresponding papers. For multi-feature aggregation methods, to achieve a fair comparison, we make two efforts to improve the results of the multi-feature methods: firstly, we utilize the same video features; secondly, we adopt Bi-HNRL in training. Note that, we also directly report the results of JEMC [25] since their method is based on an ensemble of several models, and it is very difficult to exactly re-implement the details. Table 1 and Table 2 show the overall performance of MFGATN and all the baselines on MSR-VTT and MSVD datasets, respectively. The experimental results reveal a number of interesting points:

- The performance of multi-feature aggregation methods are obviously better than that of single-feature methods, which proves the significance of utilizing complementary cues from videos to improve the video-text retrieval. For instance, the simple concatentation method achieves 3.2%, 2.7%, and 3.8% relative improvements compared with the prior state-of-the-art single-feature HGR method in R@1, R@5, and R@10 on MSR-VTT dataset, respectively.

- The proposed MFGATN approach achieves 24.7%, 14.6%, and 11.3% relative improvements compared with the prior state-of-the-art MoEE method in R@1, R@5, and R@10 on the MSR-VTT dataset, respectively. The major reason is that MoEE obtains the overall similarity by the weighted sum of each expert's similarity, but handling each modality independently, which inevitably achieve unsatisfactory results.
- The proposed MFGATN approach achieves 21%, 14.2%, and 11.9% relative improvements compared with the prior state-of-the-art CE method in R@1, R@5, and R@10 on the MSR-VTT dataset, respectively. Similarly, MFGATN achieves 17.6%, 16.7%, and 10.7% relative improvements compared with the CE method in R@1, R@5, and R@10 on the MSVD dataset, respectively. This is due to that the CE method only strengthens (or weakens) some dimensions of the input signal. Therefore, it is not able to capture high-level inter-modality information among multiple features present in videos.

In a nutshell, the multi-feature aggregation methods outperform the previous state-of-the-art single-feature methods, which demonstrates that multiple features can help to boost the performance of complicated video retrieval. Furthermore, MFGATN shows significant superiority over other multi-feature aggregation methods, indicating the benefit of high-level semantics information interchange among multiple features.

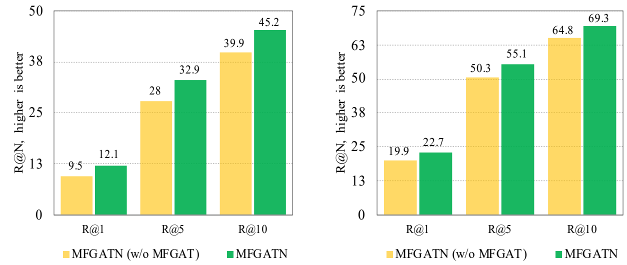
4.4 Componential Analysis

In this subsection, we present an ablation study to explore how the performance of the proposed method is affected by different components, including the multi-feature graph attention module and different loss functions in training.

Effectiveness of the multi-feature graph attention module: In order to further explore the effectiveness of the proposed multi-feature graph attention module, we devise an ablation study on MFGATN (w/o. MFGAT). To be specific, MFGATN (w/o. MFGAT) is the variant of MFGATN method which removes the MFGAT module from the full MFGATN. We see that our propose MFGATN (full) method achieves 27.4%, 17.5%, and 13.2% relative improvements compared with MFGATN (w/o. MFGAT) method in R@1, R@5, and R@10 on the MSR-VTT dataset, respectively (see Figure 4(a)). Similarly, MFGATN (full) method achieves 14.1%, 9.5%, and 6.9% relative improvements compared with MFGATN (w/o. MFGAT) method in R@1, R@5, and R@10 on the MSVD dataset, respectively (see Figure 4 (b)). The proposed MFGATN (full) method achieves significant improvements compared with MFGATN (w/o. MFGAT), which indicates that the MFGAT module plays an essential role in the video-text retrieval task.

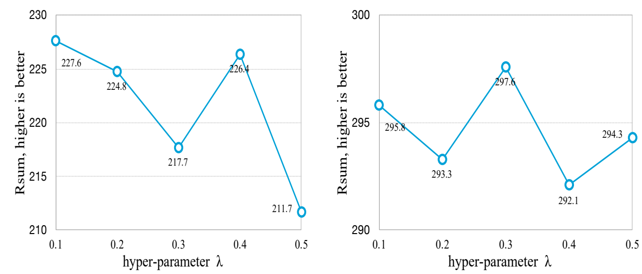
Loss functions: We compare our proposed dual constraint ranking loss with existing ranking loss function to verify the effectiveness. For instance, (1) Bi-direction hard-negatives ranking loss function (Bi-HNRL), which only considers the inter-modal ranking constraint. (2)Dual Constraint Ranking Loss (DCRL), which simultaneously considers the inter-modal ranking constraint and the intra-modal structure constraint. The results, presented in Table 3 and Table 4, demonstrate the contribution of simultaneously consideration of inter-modal ranking constraint and the intra-modal structure constraint. Specifically, our propose DCRL achieves 2.5%, 2.4%, and 2.3% relative improvements compared with Bi-HNRL in R@1, R@5, and R@10 on the MSR-VTT dataset, respectively. Similarly, DCRL achieves 2.7%, 3.8%, and 2.97% relative improvements compared with Bi-HNRL in R@1, R@5, and R@10 on the MSVD dataset, respectively.

Moreover, λ is a crucial hyper-parameter to balance the inter-modal ranking constraint and the intra-modal structure constraint. Therefore, we further explore the effect of this hyper-parameter by varying it from 0.1 to 0.5. Figure 5 shows the impact of this hyper-parameter on the MSR-VTT and MSVD datasets. From the Figure. 5, we note that the MSR-VTTdataset achieves the best performance while the λ is set to 0.1, and MSVD dataset achieves the best performance while the λ is set to 0.3.



(a) MSR-VTT Dataset **(b) MSVD Dataset**

Figure 4: Performance Evaluation Results of Ablation Model.



(a) MSR-VTT Dataset **(b) MSVD Dataset**

Figure 5: Effect of the hyperparameter λ for the intra-modal structure constraint.

4.5 Qualitative Analyses

To qualitatively validate the effectiveness of the MFGATN method, we visually present several results of the multi-feature aggregation methods. Figure. 6 shows the result results of MFGATN, CE, and MoEE on the MSR-VTT dataset, respectively. For each method, we show frames from the top-5 ranked videos (the ground truth video is indicated by a red box). Moreover, we also report the GT rank metric, which is the ground-truth rank of the relevant video returned by models. Higher ranks indicate better performance.

Table 3: Results of MFGATN with the different loss functions on the MSR-VTT dataset.

Method	Text-to-Video Retrieval				Video-to-Text Retrieval				rsum
	R@1	R@5	R@10	Med R	R@1	R@5	R@10	Med R	
Bi-HNRL	11.8	32.1	44.2	14	20.7	49.5	63.2	6	221.5
DCRL ($\lambda = 0.1$)	**12.1**	**32.9**	**45.2**	**13**	**21.4**	**51.2**	**64.8**	**5**	**227.6**

Table 4: Results of MFGATN with the different loss functions on the MSVD dataset.

Method	Text-to-Video Retrieval				Video-to-Text Retrieval				rsum
	R@1	R@5	R@10	Med R	R@1	R@5	R@10	Med R	
Bi-HNRL	22.1	53.1	67.3	5	25.5	55.8	66.4	4	290.2
DCRL ($\lambda = 0.3$)	**22.7**	**55.1**	**69.3**	**4**	**27.8**	**55.8**	**66.9**	**4**	**297.6**

Text Query: There is a woman surfing on the powerful waves.

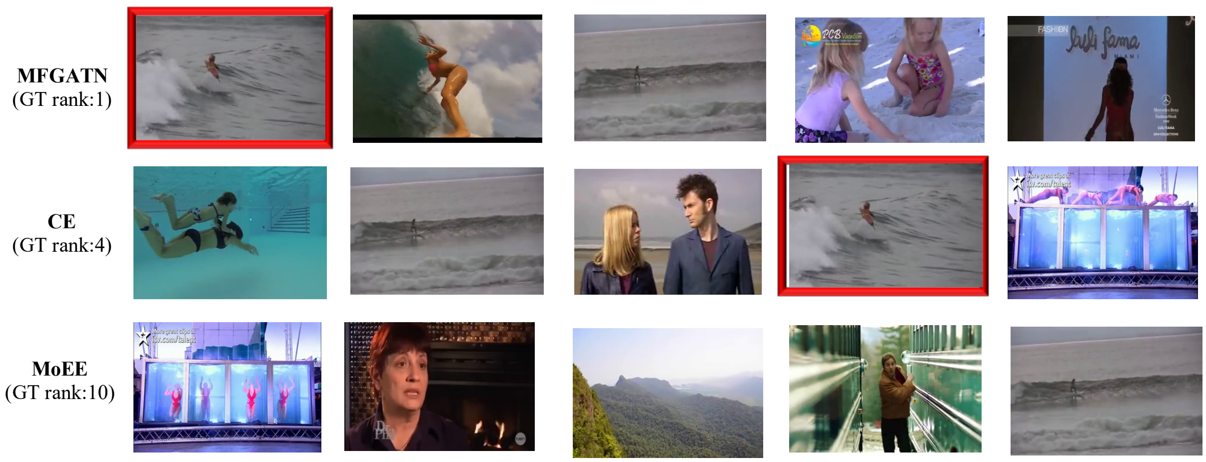


Figure 6: Visualizations of the text-to-video retrieval results on the MSR-VTT dataset. We visualize the top-5 ranked videos given the same query, and the ground truth video is indicated by a red box.

As illustrated in Figure 6, we adopt 'There is woman surfing on the powerful waves' as the same query text for all the methods. In the top line, the proposed MFGATN model successfully retrieves the correct video given the query text and the GT rank is 1, which visually demonstrates the suporiority of our method. In the middle line, CE also retrieves the correct video given the query text. However, the GT rank is 4, indicating the lack of the interchange of high-level semantic information among multiple features indeed degrades the performance. In the bottom line, MoEE fails to retrieve the correct video among the top-5 returned videos, which shows the deficiency of its simple concatenation. In conclusion, our method can retrieve more accurately than other multi-feature aggregation methods.

What's more, we can also observe that the top-ranked videos in MFGATN are more reasonable. The mismatched videos retrieved by MFGATN contain partial content with the query text (i.e., woman or waves), and are similar to the matched video, while CE and MoEE make quite different mismatched video rank higher. Providing reasonable candidates is meaningful in the real retrieval scenario where users are not very confident with their input queries. Our MFGATN can achieve this with the intra-modal structure constraint.

5 CONCLUSION

In this paper, we have proposed a novel Multi-Feature Graph Attention Network for video-text retrieval. Specifically, we introduce a multi-feature graph attention module to aggregate the multiple features in videos, and design a Dual Constraint Ranking Loss to consider both the inter-modal ranking constraint and the intra-modal structure constraint. Experiments on the MSR-VTT and the MSVD datasets have demonstrated the significant improvements of our method. In future work, we will explore the performance of MFGATN on other video understanding tasks such as clustering and summarisation.

6 ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grants 62006221, the National Key Research and Development Program of China (No. 2018YFC0825102, No.2019YFC0850202),the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China, China (No. SKLMCC2020KF004), the Key Research Program of Frontier Sciences, CAS, Grant NO ZDBS-LY-7024.

REFERENCES

[1] Samuel Albanie, Yang Liu, Arsha Nagrani, Antoine Miech, Ernesto Coto, Ivan Laptev, Rahul Sukthankar, Bernard Ghanem, Andrew Zisserman, Valentin Gabeur, Chen Sun, Karteek Alahari, Cordelia Schmid, Shizhe Chen, Yida Zhao, Qin Jin, Kaixu Cui, Hui Liu, Chen Wang, Yudong Jiang, and Xiaoshuai Hao. 2020. The End-of-End-to-End: A Video Understanding Pentathlon Challenge (2020). *CoRR* abs/2008.00744 (2020).

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5297–5307.

[3] David L Chen and William B Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Annual Meeting of the Association for Computational Linguistics*. 190–200.

[4] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[5] Jianfeng Dong, Xirong Li, and Cees G M Snoek. 2016. Word2VisualVec: Image and Video to Sentence Matching by Visual Feature Prediction. *arXiv:1604.06838* (2016).

[6] 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* (2018), 3377–3388.

[7] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual Encoding for Zero-Example Video Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9346–9355.

[8] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference*.

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marcaurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Neural Information Processing Systems*. 2121–2129.

[10] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. 2018. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7181–7189.

[11] Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, Weiping Wang, and Dan Meng. 2021. What Matters: Attentive and Relational Feature Aggregation Network for Video-Text Retrieval. In *IEEE International Conference on Multimedia and Expo*.

[12] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. 2018. Learning Semantic Concepts and Order for Image and Sentence Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6163–6171.

[13] Unmesh Joshi and Jacopo Urbani. 2020. Searching for Embeddings in a Haystack: Link Prediction on Knowledge Graphs with Subgraph Pruning. In *International World Wide Web Conferences*. 2817–2823.

[14] Ryan Kiros, Ruslan Salakhutdinov, and Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In *Neural Information Processing Systems*.

[15] Tetsunori Kobayashi. 2016. Improving semantic video indexing: Efforts in Waseda TRECVID 2015 SIN system. In *International Conference on Acoustics, Speech and Signal Processing*.

[16] Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, and Bo Zheng. 2020. Adversarial Multimodal Representation Learning for Click-Through Rate Prediction. In *International World Wide Web Conferences*. 827–836.

[17] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. 2019. W2VV++: Fully Deep Learning for Ad-hoc Video Search. In *ACM Multimedia*. 1786–1794.

[18] Wang Liang. 2017. Learning Semantic Concepts and Order for Image and Sentence Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[19] Guiliang Liu, Xu Li, Jiakang Wang, Mingming Sun, and Ping Li. 2020. Extracting Knowledge from Web Text with Monte Carlo Tree Search. In *International World Wide Web Conferences*. 2585–2591.

[20] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. 2019. Use What You Have: Video retrieval using representations from collaborative experts. In *British Machine Vision Conference*.

[21] Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. 2017. Query and Keyframe Representations for Ad-hoc Video Search. In *ACM International Conference on Multimedia Retrieval*. 407–411.

[22] Antoine Miech, Ivan Laptev, and Josef Sivic. 2017. Learnable pooling with Context Gating for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[23] Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a Text-Video Embedding from Incomplete and Heterogeneous Data. *arXiv:1804.02516* (2018).

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations*.

[25] 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 *ACM International Conference on Multimedia Retrieval*. 19–27.

[26] Hyeonseob Nam, Jungwoo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2156–2164.

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[28] Yale Song and Mohammad Soleymani. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1979–1988.

[29] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2017. Order-Embeddings of Images and Language. In *International Conference on Learning Representations*.

[30] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5005–5013.

[31] Dayan Wu, Qi Dai, Jing Liu, Bo Li, and Weiping Wang. 2019. Deep Incremental Hashing Network for Efficient Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9069–9077.

[32] Dayan Wu, Zheng Lin, Bo Li, Mingzhen Ye, and Weiping Wang. 2017. Deep Supervised Hashing for Multi-Label and Large-Scale Image Retrieval. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. 150–158.

[33] Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. 2014. Zero-Shot Event Detection Using Multi-modal Fusion of Weakly Supervised Concepts. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2665–2672.

[34] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5288–5296.

[35] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. 2015. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Association for the Advancement of Artificial Intelligence*. 2346–2352.

[36] Dejie Yang, Dayan Wu, Wanqian Zhang, Haisu Zhang, Bo Li, and Weiping Wang. 2020. Deep Semantic-Alignment Hashing for Unsupervised Cross-Modal Retrieval. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. 44–52.

[37] Naokazu Yokoya. 2016. Learning Joint Representations of Videos and Sentences with Web Image Search. In *European Conference on Computer Vision*.

[38] Jin Yuan, Zhengjun Zha, Yaotao Zheng, Meng Wang, Xiangdong Zhou, and Tatseng Chua. 2011. Learning concept bundles for video search with complex queries. In *ACM Multimedia*. 453–462.

[39] Jin Yuan, Zhengjun Zha, Yantao Zheng, Meng Wang, Xiangdong Zhou, and Tatseng Chua. 2011. Utilizing Related Samples to Enhance Interactive Concept-Based Video Search. *IEEE Transactions on Multimedia* (2011), 1343–1355.

[40] Wanqian Zhang, Dayan Wu, Jing Liu, Bo Li, Xiaoyan Gu, Weiping Wang, and Dan Meng. 2019. Fast and Multilevel Semantic-Preserving Discrete Hashing. In *British Machine Vision Conference*. 157.

[41] Wanqian Zhang, Dayan Wu, Yu Zhou, Bo Li, Weiping Wang, and Dan Meng. 2020. Binary Neural Network Hashing for Image Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[42] Wanqian Zhang, Dayan Wu, Yu Zhou, Bo Li, Weiping Wang, and Dan Meng. 2020. Deep Unsupervised Hybrid-similarity Hadamard Hashing. In *Proceedings of the ACM International Conference on Multimedia*. 3274–3282.

[43] Shu Zhao, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, and Weiping Wang. 2020. Asymmetric Deep Hashing for Efficient Hash Code Compression. In *ACM International Conference on Multimedia*. 763–771.