

# Multi-Modal Cross-Domain Alignment Network for Video Moment Retrieval

Xiang Fang, Daizong Liu *Student Member, IEEE*, Pan Zhou, *Senior Member, IEEE*, YuChong Hu *Member, IEEE*

**Abstract**—As an increasingly popular task in multimedia information retrieval, video moment retrieval (VMR) aims to localize the target moment from an untrimmed video according to a given language query. Most previous methods depend heavily on numerous manual annotations (*i.e.*, moment boundaries), which are extremely expensive to acquire in practice. In addition, due to the domain gap between different datasets, directly applying these pre-trained models to an unseen domain leads to a significant performance drop. In this paper, we focus on a novel task: cross-domain VMR, where fully-annotated datasets are available in one domain (“source domain”), but the domain of interest (“target domain”) only contains unannotated datasets. As far as we know, we present the first study on cross-domain VMR. To address this new task, we propose a novel Multi-Modal Cross-Domain Alignment (MMCDA) network to transfer the annotation knowledge from the source domain to the target domain. However, due to the domain discrepancy between the source and target domains and the semantic gap between videos and queries, directly applying trained models to the target domain generally leads to a performance drop. To solve this problem, we develop three novel modules: (i) a domain alignment module is designed to align the feature distributions between different domains of each modality; (ii) a cross-modal alignment module aims to map both video and query features into a joint embedding space and to align the feature distributions between different modalities in the target domain; (iii) a specific alignment module tries to obtain the fine-grained similarity between a specific frame and the given query for optimal localization. By jointly training these three modules, our MMCDA can learn domain-invariant and semantic-aligned cross-modal representations. Extensive experiments on three challenging benchmarks (ActivityNet Captions, Charades-STA and TACoS) illustrate that our cross-domain method MMCDA outperforms all state-of-the-art single-domain methods. Impressively, MMCDA raises the performance by more than 7% in representative cases, which demonstrates its effectiveness.

## I. INTRODUCTION

As an important yet challenging multimedia task, video moment retrieval (VMR) [1]–[9] has received increasing at-

The first two authors contributed equally to the paper writing and experiments. Corresponding author: Pan Zhou.

Xiang Fang is with the Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering Huazhong University of Science and Technology, Wuhan 430074, China (E-mail: xfang9508@gmail.com).

Daizong Liu is with Wangxuan Institute of Computer Technology, Peking University, No. 128, Zhongguancun North Street, Beijing, China (E-mail: dzliu@stu.pku.edu.cn).

Pan Zhou is with the Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China (E-mail: panzhou@hust.edu.cn).

Yuchong Hu is with the School of Computer Science and Technology, Key Laboratory of Information Storage System Ministry of Education of China, Huazhong University of Science and Technology, Wuhan 430074, China (E-mail: yuchonghu@hust.edu.cn).

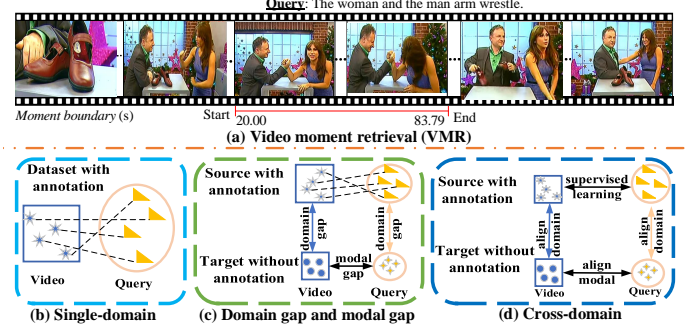


Fig. 1. (a) An example of the video moment retrieval (VMR) task. In an annotated dataset, given a language query, VMR aims to retrieve and rank videos frames based on how well they match the text description. (b) Single-domain models (previous VMR methods) rely on numerous annotations to acquire the annotation knowledge and the semantic relationship between videos and queries. These single-domain models are trained and tested on the same domain (the same dataset). (c) Due to the domain gap (domain discrepancy) and the modal gap (semantic gap), we cannot directly apply the model trained in the source domain to the target domain. (d) Our cross-domain model aims to transfer the annotation knowledge from the source domain to the target domain by aligning different domains and modalities.

tention due to its wide potential applications, such as human-computer interaction [10]–[14] and video understanding [15], [15]–[24]. Given a sentence query, VMR aims to ground the most relevant moment from an untrimmed video. As shown in Fig. 1(a), most of the contents in the video are irrelevant to the query while only a short moment matches it. A well-designed model requires grounding accurate start and end timestamps of the target moment. In practice, VMR is an extremely challenging task because the desired model should (i) cover various moment lengths in multiple scenarios; (ii) bridge the semantic gap between different modalities (video and query); (iii) understand the semantic details of different modalities to extract modal-invariant features for optimal retrieval.

Most previous VMR methods [4], [14], [25]–[30] refer to a fully-supervised setting where all moment-query pairs and moment boundaries are annotated in detail. For instance, some of them [31]–[37] utilize a scan-and-ranking framework that first pre-defines a set of candidate segment proposals and then calculates their matching degree with query semantics for ranking and localization; others [38]–[42] directly predict the moment boundary by a regression strategy. Although these respectable methods have achieved decent progress, they are data-hungry, relying on numerous manual annotations (moment boundaries), which limits their applications. For example, more than 300,000 unannotated videos (with 80,000 hours of video length) are uploaded to YouTube every day.

Obviously, it is not feasible to manually annotate such large-scale video data. To alleviate the reliance to a certain extent, some weakly-supervised methods [12], [43]–[50] are proposed to only utilize coarse video-level annotations for training. Unfortunately, this still requires us to understand each video and to provide the corresponding video-level annotation, which is significantly time-consuming and labor-intensive.

Although the above supervised methods have achieved great progress in some cases, they are all trained in a supervised learning manner based on manually annotated datasets, which are always expensive to collect in practice. In most multimedia applications, we always collect a few fully-annotated datasets and a massive number of unannotated datasets. Hence, we pose a brand-new task: *can we transfer the annotation knowledge from the fully-annotated source dataset to the unannotated target dataset?* We demonstrate this new task “cross-domain video moment retrieval” (cross-domain VMR) as shown in Fig. 1(d). To the best of our knowledge, there is no such task proposed in previous works for VMR. Different from previous VMR works under the single-domain setting as shown in Fig. 1(b), our cross-domain model first gains the annotation knowledge from the source domain, and then transfers the knowledge into the target domain as shown in Fig. 1(d). Due to the domain discrepancy and semantic gap as shown in Fig. 1(c), if previous single-domain VMR methods are directly employed to the unannotated target dataset, they will suffer from a performance drop. Therefore, cross-domain VMR is a significant and challenging task.

Although many single-modal domain adaptation works [51]–[55] have been proposed, they can only learn the feature representation of one modality. As a significant multimedia task, the VMR task requires the alignment knowledge of two modalities. Therefore, previous single-modal domain adaptation methods cannot be directly used for the VMR task due to the inability to preserve the relations between two modalities at the same time. Recently, some multi-modal domain adaptation methods have been proposed for some specific multimedia tasks, *e.g.*, fine-grained action recognition [56], and emotion recognition [57]. Someone might have a question: if we can introduce these methods to our proposed cross-domain VMR? Unfortunately, there is a large gap between VMR and other multimedia tasks. Directly introducing these works into VMR meets limitations of the more complex multi-modal scenarios and the larger domain gap in VMR datasets. Different from previous multi-modal domain adaptation tasks that focus on matching visual and textual information for a classification-based task, our proposed cross-domain VMR is more challenging and needs to transfer the query-related moment knowledge across different datasets in a more complicated retrieval task.

To solve this task, we propose a Multi-Modal Cross-Domain Alignment (MMCDA) network to learn domain-invariant and semantic-aligned cross-modal representations. As shown in Fig. 2, MMCDA contains three main parts: (i) two shared multi-modal feature encoders to extract discriminative modal-specific features, (ii) a cross-modal attention part to learn domain-agnostic video-query interaction, (iii) a multi-modal cross-domain alignment part to close the domain discrepancy and the semantic gap via three carefully-designed alignment

modules: domain alignment, cross-modal alignment, and specific alignment.

To alleviate the domain discrepancy, we first design a brand-new domain alignment module to minimize the maximum mean discrepancy (MMD) distance between different domains. Moreover, an intra-sample distribution function and an inter-sample distribution function are proposed to align the feature distributions of the source and target domains. Based on this domain alignment module, we can obtain the domain-invariant representation for each modality. However, only aligning different domains is not enough due to the existence of the semantic gap between videos and queries in the unannotated target domain. To bridge such a semantic gap, we further develop a cross-modal alignment module. In particular, we first project these multi-modal features into a joint embedding space, which aims to maximize the similarity between the video and corresponding query and to minimize the similarity between the video and non-matching query, *versa vice*. Then, we align the feature distributions of different modalities based on the intra-sample distribution function and the inter-sample distribution function. Considering that some specific frames near a moment boundary might be assigned to another moment incorrectly, we also propose a specific alignment module to improve the retrieval accuracy. Particularly, we enhance the characteristics of the specific frame by maximizing its probability belonging to the corresponding query. By integrating the aforementioned three alignment modules, our MMCDA is able to achieve significant performance on the target domain without any annotations.

Our main contributions are summarized as follows:

- We present a novel task: cross-domain VMR, which aims to transfer the annotation knowledge from the source domain to the target domain. To our best knowledge, our proposed MMCDA is the first attempt to make VMR adaptive to different domains rather than a single domain in previous VMR methods.
- Three alignments modules (domain alignment, cross-modal alignment, and specific alignment) are designed in MMCDA to learn domain-invariant and semantic-aligned cross-modal representations.
- Experimental results on three popular datasets (ActivityNet Captions, Charades-STA and TACoS) show the effectiveness of MMCDA. To our surprise, it significantly outperforms state-of-the-art single-domain VMR methods in representative cases. Extensive ablation studies illustrate the functions of each module.

## II. RELATED WORKS

**Fully-supervised video moment retrieval.** Most of the existing methods are fully supervised [8], requiring all annotated moment-query pairs and labeled moment boundaries. With adequate annotations, these methods try to align multi-modal features to predict the moment boundary from an untrimmed video according to a given query. Based on a two-stage multi-modal matching strategy, many works [31]–[37] first sample candidate moment proposals, and subsequently integrate a given query with moment representations by a matrix operation. For example, by introducing a multi-level model, Xu *et*

*al.* [33] integrates visual and textual features earlier and further re-generates queries as an auxiliary task. To enhance the video representation understanding, Chen *et al.* [58] captures the evolving fine-grained frame-by-word interactions between videos and queries. Although these methods achieve good performances, they are severely proposal-dependent and time-consuming. Without using proposals, the latest methods [38]–[42] integrate the query representation with those video moments individually, and then calculate their matching scores. By introducing the textual information as a critical prior, Yuan *et al.* [59] modulate temporal convolution operations to compose and correlate video contents. Based on a multi-step reasoning network, Liu *et al.* [36] modulate conditioned temporal relation for optimal retrieval.

**Weakly-supervised video moment retrieval.** The aforementioned fully-supervised methods are annotation-dependent. In order to alleviate the dependence to a certain extent, several weakly-supervised VMR methods [45]–[47], [49] are under a weakly-supervised setting, which only utilizes coarse video-level annotations to obtain accurate retrieval. For a weakly-supervised VMR task, Duan *et al.* [46] decompose it into two sub-tasks: event captioning and query localization. Duan *et al.* [46] first assume that each caption describes only one temporal moment, and then design a cycle network to train the model. As the pioneering work for weakly-supervised VMR, Mithun *et al.* [47] learn a joint representation between the video and the query by proposing a text-guided-attention network and utilizing an attention weight. To improve the exploration and exploitation, Lin *et al.* [49] choose the top-K proposals and measure the semantic similarity between the video and the query. By proposing Semantic Completion Network, Lin *et al.* [49] treat the masked query as input and predict the masked words from the video features.

**Unsupervised domain adaptation.** Unsupervised domain adaptation (UDA) aims to transfer the knowledge from the label-rich source domain to the unlabelled target domain, which is called *domain shift* [60]–[62]. This shift usually encounters some challenges: (i) the collected data are from diverse resources [63]; (ii) the training data in the target domain is limited [64]; (iii) the source and target domains have different modalities [65]. A general solution is to align the feature distribution between the two domains. Many UDA methods align the feature distribution by minimizing the disagreement of data distribution in different domains [66], [67], or learning an adversarial domain-classifier to find domain-invariant features [68], [69]. For example, for the activity recognition task, Song *et al.* [70] establish the cross-modal domain alignment via self-supervised contrastive framework to learn the joint clip-level and video-level representation alignment for video-based UDA. Considering that video data is usually associated with multi-modal information (RGB and optical flow), Kim *et al.* [71] propose a unified framework for video domain adaptation, which simultaneously regularizes cross-modal and cross-domain feature representations. As for the text-video retrieval task, Chen *et al.* [72] propose a new benchmark and investigate the UDA task for text-video retrieval in this setting. Based on standard backpropagation training, Ganin *et al.* [73] present a network that allows large-

scale training by massive annotated data in the source domain and a large number of unannotated data in the target domain. By introducing adversarial learning to model domain shift between different domains, Long *et al.* [74] design a deep domain adaptation hashing network for binary representation generation. Long *et al.* [75] propose a conditional adversarial domain adaptation network, which conditions the adversarial adaptation methods on discriminative information conveyed when predicting classification results. However, these methods cannot apply to our cross-domain VMR task because (i) most UDA methods only align the unimodal features, while we need to align the multi-modal features (both queries and videos); (ii) many UDA methods aim to tackle classification-based tasks (*e.g.*, activity recognition [70], [71], [76], object detection [77] and text-video retrieval [72]) rather than our complicated retrieval task. Thus, cross-domain VMR is more challenging than general UDA tasks.

### III. PROPOSED MMCD A NETWORK

Firstly, we carefully formulate our novel problem and introduce our proposed MMCD A. Then, we present our framework (feature encoder, cross-modal attention, and multi-modal cross-domain alignment). In the multi-modal cross-domain alignment, we highlight three well-designed alignment modules: domain alignment, cross-modal alignment, and specific alignment. Finally, we demonstrate our training and inference phases in detail.

#### A. Problem Formulation

In our cross-domain VMR task, we collect a fully-annotated source dataset as:

$$\mathbf{X}^s = \{ \{ \mathbf{V}_i^s \}_{i=1}^{I^s}, \{ \mathbf{Q}_j^s \}_{j=1}^{J^s}, \{ (s_j, e_j)^s \}_{j=1}^{J^s} \},$$

where  $\mathbf{V}_i^s$  is the  $i$ -th untrimmed video;  $\mathbf{Q}_j^s$  and  $(s_j, e_j)^s$  are the  $j$ -th sentence query and moment boundary, respectively;  $(s_j, e_j)$  is the start and end timestamps of the  $j$ -th moment, which corresponds to the  $j$ -th query semantically.

Also, we are given an unannotated target dataset:

$$\mathbf{X}^t = \{ \{ \mathbf{V}_k^t \}_{k=1}^{I^t}, \{ \mathbf{Q}_l^t \}_{l=1}^{J^t}, \{ (\cdot, \cdot)^t \}_{l=1}^{J^t} \},$$

where “ $(\cdot, \cdot)$ ” means that the target moment boundary is not available.  $I^s/I^t$  and  $J^s/J^t$  are the numbers of source/target videos and source/target queries, respectively<sup>1</sup>. To predict the target moment boundary  $(\cdot, \cdot)^t$ , we propose MMCD A to transfer the annotation knowledge from the source domain into the target domain.

As shown in Fig. 2, our MMCD A contains three main parts: feature encoder, cross-modal attention, and multi-modal cross-domain alignment. The feature encoder aims to extract the discriminative features ( $\mathbf{V}_i^s/\mathbf{V}_k^t$  and  $\mathbf{Q}_j^s/\mathbf{Q}_l^t$ ) with modal-specific information. The cross-modal attention tries to learn the domain-agnostic video-query interaction by integrating both query-to-video and video-to-query information as the bidirectional attention ( $\mathbf{V}_f^s/\mathbf{V}_f^t$ ). The multi-modal cross-domain

<sup>1</sup>In this paper, the superscript  $s$  represents the source domain, and the superscript  $t$  represents the target domain.

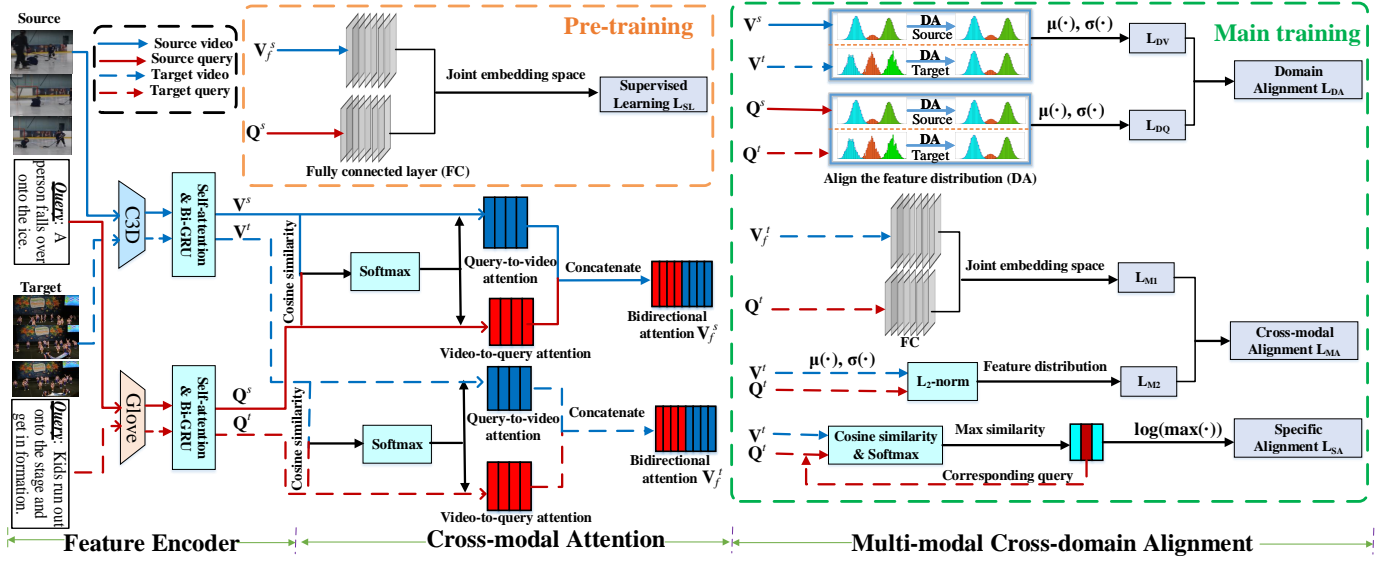


Fig. 2. A brief framework of our proposed MMCD for the cross-domain VMR task. MMCD contains three parts: the shared feature encoders to extract the video and query features in each domain, the cross-modal attention to capture the video-query relations, and the multi-modal cross-domain alignment to close the domain discrepancy and the semantic gap. MMCD aims to transfer the alignment knowledge from the source domain to the target domain. Best viewed in color.

alignment aims to learn domain-invariant and semantic-aligned cross-modal representations through three alignment modules: domain alignment, cross-modal alignment, and specific alignment.

### B. Feature Encoder

**Video encoding.** Given a video input, following [37], we first utilize a pre-trained C3D model [78] to extract the frame-wise features. Then, we leverage a multi-head self-attention module [79] to obtain the semantic dependencies in the long video context. To incorporate the contextual information, we further exploit a bi-directional GRU (Bi-GRU) network [80]. In the source domain, we denote the encoded video representation as  $V_i^s = \{v_1^s, v_2^s, \dots, v_{T_i^s}^s\}^\top \in \mathbb{R}^{T_i^s \times d}$ , where  $T_i^s$  is the frame number of the  $i$ -th source video,  $d$  is the embedding dimension. Similarly, we leverage the same video encoder for the target domain. The extracted features can be represented as  $V_k^t = \{v_1^t, v_2^t, \dots, v_{T_k^t}^t\}^\top \in \mathbb{R}^{T_k^t \times d}$ , where  $T_k^t$  is the frame number of the  $k$ -th target video.

**Query encoding.** Given a query as input, a GloVe embedding [81] is employed to obtain the word-wise features. Similar to the video encoding, we feed these word features into another Bi-GRU network and a multi-head self-attention module to further integrate the sequential query representations. In the source domain, we denote the encoded query representation as  $Q_j^s = \{q_1^s, q_2^s, \dots, q_{N_j^s}^s\}^\top \in \mathbb{R}^{N_j^s \times d}$ , where  $N_j^s$  is the word number of the  $j$ -th source query,  $d$  is the embedding dimension. Similarly, in the target domain, we use the same query encoder to obtain the query representation  $Q_l^t = \{q_1^t, q_2^t, \dots, q_{N_l^t}^t\}^\top \in \mathbb{R}^{N_l^t \times d}$ , where  $N_l^t$  is the word number of the  $l$ -th target query.

### C. Cross-Modal Attention

After encoding both videos and queries, we learn domain-agnostic video-query interactions through two attention mechanisms, shown in Fig. 2.

Firstly, we compute two cosine similarity matrices: (i)  $C^s(V_i^s, Q_j^s) \in \mathbb{R}^{T_i^s \times N_j^s}$  between  $V_i^s$  and  $Q_j^s$  in the source domain, (ii)  $C^t(V_k^t, Q_l^t) \in \mathbb{R}^{T_k^t \times N_l^t}$  between  $V_k^t$  and  $Q_l^t$  in the target domain:

$$C_{i,j}^s = \frac{\|v_i^s\|_2 \|q_j^s\|_2}{\|v_i^s\|_2 \|q_j^s\|_2}, C_{k,l}^t = \frac{\|v_k^t\|_2 \|q_l^t\|_2}{\|v_k^t\|_2 \|q_l^t\|_2}, \quad (1)$$

where  $Q_j^s/Q_l^t$  is the corresponding query semantically-matching video  $V_i^s/V_k^t$ ,  $(i, j)$  and  $(k, l)$  is the coordinates of  $C^s$  and  $C^t$ ;  $v_i^s, v_k^t$  and  $q_j^s, q_l^t$  are the frames and words in source and target domains, respectively. Specifically,  $C_r^s$  and  $C_c^s$  mean the row-wise and column-wise normalizations of  $C^s(V_i^s, Q_j^s)$  by the Softmax function, respectively.

Then, in the source domain, we design the bidirectional attention module as follows:

$$\mathcal{X}_i^s = C_r^s Q_i^s, \mathcal{Y}_i^s = C_c^s C_r^{s\top} V_i^s.$$

Similarly, in the target domain, the bidirectional attention module is as follows:

$$\mathcal{X}_k^t = C_r^t Q_k^t, \mathcal{Y}_k^t = C_c^t C_r^{t\top} V_k^t,$$

where  $\mathcal{X}_i^s \in \mathbb{R}^{T_i^s \times d}$  and  $\mathcal{X}_k^t \in \mathbb{R}^{T_k^t \times d}$  are the video-to-query attention weights;  $\mathcal{Y}_i^s \in \mathbb{R}^{T_i^s \times d}$  and  $\mathcal{Y}_k^t \in \mathbb{R}^{T_k^t \times d}$  are the query-to-video attention weights.

Finally, we compose the fused frame-wise feature sequence with the output of the bidirectional attention:

$$\begin{aligned} \{V_f^s\}_i &= \text{Bi-GRU}([V_i^s; \mathcal{X}_i^s; V_i^s \odot \mathcal{X}_i^s; V_i^s \odot \mathcal{Y}_i^s]), \\ \{V_f^t\}_k &= \text{Bi-GRU}([V_k^t; \mathcal{X}_k^t; V_k^t \odot \mathcal{X}_k^t; V_k^t \odot \mathcal{Y}_k^t]), \end{aligned} \quad (2)$$

where  $\{V_f^s\}_i \in \mathbb{R}^{T_i^s \times d}$ ;  $\{V_f^t\}_k \in \mathbb{R}^{T_k^t \times d}$ ,  $\odot$  is the element-wise multiplication operation.

#### D. Multi-Modal Cross-Domain Alignment

So far, we encode videos and queries ( $\mathbf{V}_i^s, \mathbf{Q}_j^s, \mathbf{V}_k^t, \mathbf{Q}_l^t$ ) in both domains, and obtain the fused features ( $\mathbf{V}_f^s, \mathbf{V}_f^t$ ) by cross-modal interactions. However, when we transfer the localization model trained on the source domain into the target domain, we still meet the following challenges: (i) how to close the domain discrepancy of the same modality between different domains; (ii) how to close the semantic gaps of the video and query in each domain; (iii) how to learn more representative frame-wise features to perform more accurate localization. As shown in Fig. 2, we address the above three challenges by proposing three alignment modules: (i) a domain alignment (DA) module to align the feature distributions between different domains of each modality; (ii) a cross-modal alignment (MA) module to map both video and query features into a joint embedding space and to align the feature distributions between different modalities in the target domain; (iii) a specific alignment (SA) module to enhance the characteristics of a specific frame by maximizing its probability belonging to the corresponding query.

**Domain alignment.** In our cross-domain VMR task, for each modality (video or query), the domain discrepancy is mainly due to the gap of feature distributions across different domains. To alleviate the domain discrepancy, we propose a domain alignment module for relieving the divergence of domain statistics. The basic idea is that the moment calculated on different domains are the same if the distributions of source and target domains are identical. To alleviate the domain discrepancy, we aim to shift the feature distributions in different domains as close as possible for each modality. In our task, the feature distribution consists of two folds: intra- and inter-sample distributions<sup>2</sup>. Therefore, we propose two domain alignment loss functions  $\mathcal{L}_{DV}$  and  $\mathcal{L}_{DQ}$  as follows:

$$\begin{aligned}\mathcal{L}_{DV} &= \sum_{\mathbf{V}_i^s, \mathbf{V}_k^t} MMD(\mu(\mathbf{V}_i^s), \mu(\mathbf{V}_k^t)) + MMD(\sigma(\mathbf{V}^s), \sigma(\mathbf{V}^t)), \\ \mathcal{L}_{DQ} &= \sum_{\mathbf{Q}_j^s, \mathbf{Q}_l^t} MMD(\mu(\mathbf{Q}_j^s), \mu(\mathbf{Q}_l^t)) + MMD(\sigma(\mathbf{Q}^s), \sigma(\mathbf{Q}^t)),\end{aligned}$$

where  $\mathcal{L}_{DV}$  is the video-based domain alignment loss;  $\mathcal{L}_{DQ}$  is the query-based domain alignment loss;  $\mu(\cdot)$  and  $\sigma(\cdot)$  are the intra-sample distribution function and the inter-sample distribution function respectively, which are defined by:

$$\begin{aligned}\mu(\mathbf{V}_i^s) &= \frac{1}{T_i^s} \sum_{a=1}^{T_i^s} \mathbf{v}_a^s, \sigma(\mathbf{V}^s) = \sqrt{\frac{1}{I^s} \sum_{i=1}^{I^s} (\mu(\mathbf{V}_i^s) - \frac{\sum_{i=1}^{I^s} \mu(\mathbf{V}_i^s)}{I^s})^2}, \\ \mu(\mathbf{Q}_j^s) &= \frac{1}{N_j^s} \sum_{b=1}^{N_j^s} \mathbf{q}_b^s, \sigma(\mathbf{Q}^s) = \sqrt{\frac{1}{J^s} \sum_{j=1}^{J^s} (\mu(\mathbf{Q}_j^s) - \frac{\sum_{j=1}^{J^s} \mu(\mathbf{Q}_j^s)}{J^s})^2}, \\ \mu(\mathbf{V}_k^t) &= \frac{1}{T_k^t} \sum_{c=1}^{T_k^t} \mathbf{v}_c^t, \sigma(\mathbf{V}^t) = \sqrt{\frac{1}{I^t} \sum_{k=1}^{I^t} (\mu(\mathbf{V}_k^t) - \frac{\sum_{k=1}^{I^t} \mu(\mathbf{V}_k^t)}{I^t})^2}, \\ \mu(\mathbf{Q}_l^t) &= \frac{1}{N_l^t} \sum_{d=1}^{N_l^t} \mathbf{q}_d^t, \sigma(\mathbf{Q}^t) = \sqrt{\frac{1}{J^t} \sum_{l=1}^{J^t} (\mu(\mathbf{Q}_l^t) - \frac{\sum_{l=1}^{J^t} \mu(\mathbf{Q}_l^t)}{J^t})^2},\end{aligned}$$

where  $\mu(\mathbf{V}_i^s)$  (or  $\mu(\mathbf{V}_k^t)$ ) means the intra-video distribution, which is learned by averaging these frame features in the  $i$ -th source video (or the  $k$ -th target video);  $\sigma(\mathbf{V}^s)$  (or  $\sigma(\mathbf{V}^t)$ )

means the inter-video distribution, which is obtained by computing the standard deviation of different video features in the source domain (or the target domain). Similarly,  $\mu(\mathbf{Q}_j^s)$  (or  $\mu(\mathbf{Q}_l^t)$ ) means the intra-query distribution;  $\sigma(\mathbf{Q}^s)$  (or  $\sigma(\mathbf{Q}^t)$ ) means the inter-query distribution.

In  $\mathcal{L}_{DV}$  and  $\mathcal{L}_{DQ}$ ,  $MMD(\cdot)$  is a function to restrict two sets of variables ( $\mathbf{U}, \mathbf{W}$ ) matching with each other:

$$\begin{aligned}MMD(\mathbf{U}, \mathbf{W}) &= \frac{1}{N_u^2} \sum_{j=1}^{N_u} \sum_{i=1}^{N_u} \phi(\mathbf{u}_i, \mathbf{u}_j) + \frac{1}{N_w^2} \sum_{j=1}^{N_w} \sum_{i=1}^{N_w} \phi(\mathbf{w}_i, \mathbf{w}_j) \\ &+ \frac{1}{N_u N_w} \sum_{i=1}^{N_u} \sum_{j=1}^{N_w} \phi(\mathbf{u}_i, \mathbf{w}_j),\end{aligned}\quad (3)$$

where  $\phi(\cdot, \cdot)$  means the Gaussian kernel;  $\mathbf{u}_i$  (resp.  $\mathbf{w}_j$ ) is the  $i$ -th (resp.  $j$ -th) sample from the set  $\mathbf{U}$  (resp.  $\mathbf{W}$ );  $N_u$  and  $N_w$  are the sample numbers of  $\mathbf{U}$  and  $\mathbf{W}$ , respectively.

The total domain alignment loss function is as follows:

$$\mathcal{L}_{DA} = \mathcal{L}_{DV} + \mathcal{L}_{DQ}. \quad (4)$$

Thus, by reducing the distance between two distributions based on Eq. (3), we align the feature distributions of the same modal in different domains.

**Cross-modal alignment.** Besides alleviating the domain discrepancy, we also need to bridge the semantic gap between videos and queries in the target domain, which is the key to the standard VMR task. Thus, we develop a cross-modal alignment module by mapping videos and queries into a joint embedding space and aligning the feature distributions of different modalities.

On the one hand, we introduce two projection matrices  $\mathbf{P}_V \in \mathbb{R}^{d \times d}$  and  $\mathbf{P}_Q \in \mathbb{R}^{d \times d}$  to project the video and query features into their joint embedding spaces as  $\{\mathbf{V}_p^t\}_k = \{\mathbf{V}_f^t\}_k \mathbf{P}_V$  and  $\{\mathbf{Q}_p^t\}_l = \{\mathbf{Q}_f^t\}_l \mathbf{P}_Q$ , respectively. Based on the joint embedding, the cross-modal consistent loss is formulated as:

$$\begin{aligned}\mathcal{L}_{M1} &= \mathcal{L}_{triplet}(\{\mathbf{Q}_p^t\}_l, \{\mathbf{V}_p^t\}_{l+}, \{\mathbf{V}_p^t\}_{l-}, m^t) \\ &+ \mathcal{L}_{triplet}(\{\mathbf{V}_p^t\}_k, \{\mathbf{Q}_p^t\}_{k+}, \{\mathbf{Q}_p^t\}_{k-}, m^t),\end{aligned}\quad (5)$$

where  $m^t$  is a margin in the target domain, the triplet ( $\{\mathbf{Q}_p^t\}_l, \{\mathbf{V}_p^t\}_{l+}, \{\mathbf{V}_p^t\}_{l-}, m^t$ ) are matching video  $\{\mathbf{V}_p^t\}_{l+}$  and non-matching video  $\{\mathbf{V}_p^t\}_{l-}$  to query  $\{\mathbf{Q}_p^t\}_l$ ; similarly, triplet ( $\{\mathbf{V}_p^t\}_k, \{\mathbf{Q}_p^t\}_{k+}, \{\mathbf{Q}_p^t\}_{k-}, m^t$ ) are video  $\{\mathbf{V}_p^t\}_k$  with matching query  $\{\mathbf{Q}_p^t\}_{k+}$  and non-matching query  $\{\mathbf{Q}_p^t\}_{k-}$ ;  $\mathcal{L}_{triplet}$  is a triplet loss defined as follows:

$$\begin{aligned}\mathcal{L}_{triplet}(\mathbf{A}, \mathbf{P}, \mathbf{N}, m) &= \sum_{\mathbf{A}, \mathbf{P}} \sum_{\mathbf{N}} \max(m + \|\mathbf{C}(\mathbf{A}, \mathbf{P})\|_2 \\ &- \|\mathbf{C}(\mathbf{A}, \mathbf{N})\|_2),\end{aligned}\quad (6)$$

where  $\|\cdot\|_2$  means L2-norm,  $\mathbf{A}$  is an anchor input,  $\mathbf{P}$  is a positive input matching  $\mathbf{A}$ ,  $\mathbf{N}$  is a negative input that does not match  $\mathbf{A}$ ,  $m$  is a certain margin between positive pair  $(\mathbf{A}, \mathbf{P})$  and negative pair  $(\mathbf{A}, \mathbf{N})$ . All these matching/non-matching samples are selected by pairwise ranking and the hardest negative mining strategies [82]. By minimizing Eq. (5), the similarity between a video feature and the corresponding query feature is maximized, and the similarity between all other

<sup>2</sup>A sample means a video or a query.



non-matching ones is minimized, which aligns the textual semantics and the matching visual semantics in the target domain.

On the other hand, we align the feature distributions of videos and queries. Based on the intra-sample distribution function  $\mu(\cdot)$  and the inter-sample distribution function  $\sigma(\cdot)$ , we develop the following cross-modal feature distribution loss:

$$\mathcal{L}_{M2} = \sum_{V_k^t, Q_l^t} \|\mu(V_k^t) - \mu(Q_l^t)\|_2^2 + \|\sigma(V^t) - \sigma(Q^t)\|_2^2. \quad (7)$$

Based on Eq. (7), we can make the feature distributions of queries and matching videos as close as possible.

Combining Eq. (5) and Eq. (7), we can bridge the semantic gap between videos and queries by the following cross-modal alignment loss:

$$\mathcal{L}_{MA} = \mathcal{L}_{M1} + \mathcal{L}_{M2}. \quad (8)$$

**Specific alignment.** Due to the lack of annotations in the target domain, the learned frame-wise representations might be not optimally discriminative. To avoid some frames near the target moment boundary being localized in incorrect boundaries, we develop a specific alignment module to learn more fine-grained and representative features of each specific frame by correlating to the most relevant query information.

Given a set of specific frame features  $\{v_c^t\}_{c=1}^{T_k^t}$  and a set of query features  $\{Q_l^t\}_{l=1}^{N_l^t}$ , we first calculate their cosine similarity  $C^t(v_c^t, Q_l^t)$  as follows:

$$C^t(v_c^t, Q_l^t) = \frac{\|v_c^{t\top} Q_l^t\|_2}{\|v_c^t\|_2 \|Q_l^t\|_2}. \quad (9)$$

To obtain the probability that  $v_c^t$  corresponds to  $Q_l^t$ , we regularize the similarity by the Softmax function:

$$s_{cl}^t = \text{Softmax}(C^t(v_c^t, Q_l^t)). \quad (10)$$

Then, we choose the query with the greatest probability as the correct query corresponding to  $v_c^t$ . Finally, we employ the specific alignment loss as follows:

$$\mathcal{L}_{SA} = - \sum_c \log(\max(s_{cl}^t)_{l=1}^{N_l^t}). \quad (11)$$

By minimizing Eq. (11), we can maximize the similarity between the specific frame  $v_c^t$  and the corresponding query  $Q_l^t$ . In this way, we can enhance the characteristics of the specific frame by maximizing its probability belonging to the corresponding query.

### E. Training and Testing

**Pre-training for source domain.** Since we have adequate annotations in the source domain, we first train our MMCDA network in the source domain under a fully-supervised setting. Given  $V_0^s$  and  $Q^s$ , we introduce two projection matrices  $P_v \in \mathbb{R}^{d \times T^s}$  and  $P_q \in \mathbb{R}^{d \times N^s}$  to project the video and query features into the joint semantic representation. On the joint representation, the projection for the video feature is derived as  $V_p^s = P_v V_0^s$  and the projection for the query feature is derived as  $Q_p^s = P_q Q^s$ . Similar to the joint embedding in

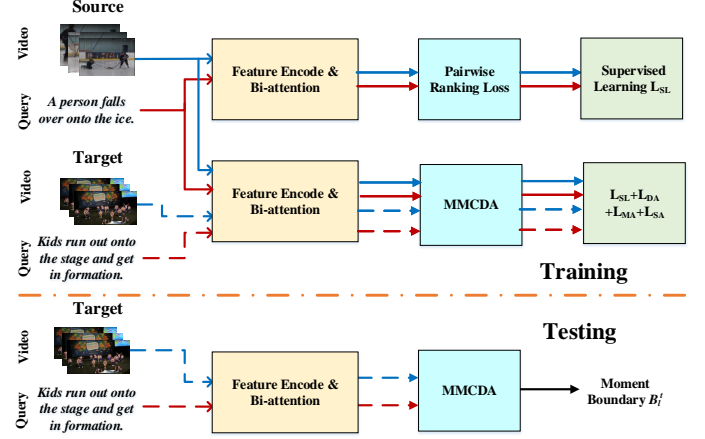


Fig. 3. The training and testing process of our MMCDA.

the cross-domain alignment (Eq. (5)), we utilize the pairwise ranking loss with the hardest negative samples as follows:

$$\mathcal{L}_{SL} = \mathcal{L}_{triplet}(\{Q_p^s\}_j, \{V_p^s\}_{j+}, \{V_p^s\}_{j-}, m^s) + \mathcal{L}_{triplet}(\{V_p^s\}_i, \{Q_p^s\}_{i+}, \{Q_p^s\}_{i-}, m^s), \quad (12)$$

where  $m^s$  is the certain margin in the source domain. Eq. (12) maximizes the similarity between an embedded video feature and the corresponding query feature, and minimizes similarity to all other non-matching ones. In Eq. (12), the first term ensures that given a query input, the matching video should have a larger similarity than the non-matching videos. As for its second term, when we have a video input, the matching query inputs should have a larger similarity than the non-matching query inputs.

**Main training.** For the target domain, we do not have the annotated temporal boundaries of each moment-query pair. To transfer the annotation knowledge from the source domain into the target domain, we combine different alignment modules (Eq. (4), Eq. (8), Eq. (11), and Eq. (12)) into a multi-task loss function (*i.e.*, the final loss function  $\mathcal{L}_{final}$ ) as follows:

$$\mathcal{L}_{final} = \mathcal{L}_{SL} + \gamma_1 \mathcal{L}_{DA} + \gamma_2 \mathcal{L}_{MA} + \gamma_3 \mathcal{L}_{SA}, \quad (13)$$

where  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$  are parameters to balance the importance between different modules.

Note that our training procedure is divided into two stages shown in Fig. 3. At the first stage, we only use  $\mathcal{L}_{SL}$  for pre-training in the source domain. At the second stage, we leverage the final objective function  $\mathcal{L}_{final}$  for main training in the source and target domains. The procedure not only effectively gains the annotation knowledge from the source domain, but also transfers the knowledge to the target domain. We can learn the domain-invariant representation [62] by the domain alignment (Eq. (4)) and obtain the semantic aligned cross-modal representations by the cross-modal alignment (Eq. (8)) and the semantic alignment (Eq. (11)). Thus, based on Eq. (13), we can learn domain-invariant and semantic aligned cross-modal representations to improve the generalization of our network.

**Inference.** Following [47], for a set of frames and a given query, we take the frame with the greatest similarity as the

TABLE I

STATISTICS OF THE DATASETS, WHERE  $L(\text{VIDEO})$  IS AVERAGE LENGTH OF VIDEOS IN SECONDS,  $L(\text{QUERY})$  IS AVERAGE NUMBER OF WORDS IN SENTENCE QUERY,  $L(\text{MOMENT})$  IS AVERAGE LENGTH OF MOMENTS IN SECONDS, AND  $S(\text{MOMENT})$  IS THE STANDARD DEVIATION OF MOMENT LENGTH IN SECONDS.

Dataset	Domain	# Videos	# Annotations	$L(\text{video})$	$L(\text{query})$	$L(\text{moment})$	$S(\text{moment})$
ActivityNet Caption	Open	14,926	54,926	117.61s	14.78	36.18s	40.18s
Charades-STA	Indoors	6,672	16,128	30.59s	7.22	8.22s	3.59s
TACoS	Cooking	127	18,818	287.14s	10.05	5.45s	7.56s

moment center, and the moment boundaries are localized by comparing the scores of the adjacent frames with sudden change. In particular, the left moment boundary is localized where one's left-frame similarity is lower than the right, and vice versa.

#### IV. EXPERIMENTAL RESULTS

##### A. Datasets and Evaluation Metrics

We conduct experiments on three challenging benchmark datasets: TACoS [83], Charades-STA [84], and ActivityNet Captions [85], summarized in Table I.

**TACoS.** TACoS [83] contains 127 videos, and the average video length is about 7 minutes. These videos are mainly about cooking scenarios, thus lacking diversity. The cooking activities appear in the same kitchen scene with some slightly varied cooking objects. For a fair comparison, we follow the same splitting protocol as [84].

**Charades-STA.** Introduced for the video moment retrieval task, Charades-STA [84], [86] is mainly about daily indoor activities with 9,848 videos and 16,128 video-query pairs. Videos in the Charades dataset are mainly about indoor everyday activities. The average video length is about 0.5 minutes. The annotations are generated by query decomposition and keywords matching with a manual check. As the original Charades dataset has only temporal activity localization and video-level paragraph description annotations, labels for video moment retrieval are added in Charades-STA later.

**ActivityNet Captions.** ActivityNet Captions [85] is a large-scale dataset from open activities including 20,000 untrimmed videos with 100,000 queries from YouTube. These videos in the ActivityNet Captions dataset are open-domain and much longer with an average duration of about 2 minutes and queries are about 13.5 words.

Note that in our experiments, we use one dataset as the source domain and the rest datasets as the target domain. We abbreviate "TACoS", "Charades-STA", and "ActivityNet Captions" as "T", "C", and "A" respectively. "A→T" means that we transfer the annotation knowledge from the ActivityNet Captions dataset (source domain) to the TACoS dataset (target domain). We consider six transfer tasks: A→T, C→T, A→C, T→C, T→A, and C→A. Meanwhile, we design three fully-supervised single-domain tasks: A→A, C→C, and T→T. In each single-domain task, for our MMCDA, we only perform pre-training (rather than main training) on a dataset and testing on the same dataset.

Following [87], we use the metric " $R@n$ ,  $\text{IoU}=m$ " for evaluation. " $R@n$ ,  $\text{IoU}=m$ " calculates the ratio of at least one of top- $n$  selected moments having an intersection over

union (IoU) larger than  $m$ . The larger metric means the better performance. In our experiments, we use  $n \in \{1, 5\}$  for all datasets,  $m \in \{0.3, 0.5, 0.7\}$  for ActivityNet Captions and Charades-STA,  $m \in \{0.1, 0.3, 0.5\}$  for TACoS.

##### B. Implementation Details

For video encoding, following [36], we resize every frame into  $112 \times 112$  pixels shape to encode videos, and then utilize a pre-trained C3D model [78] to extract frame-wise features for TACoS, ActivityNet Captions, and Charades-STA datasets. The length of each video is set to 200 in TACoS and ActivityNet Captions and 64 in Charades-STA. For query encoding, each word is embedded in a 300-dimensional vector with the Glove embedding [81]. The hidden state dimension of the Bi-GRU network is set to 512. The hyperparameters  $\gamma_1$ ,  $\gamma_2$ ,  $\gamma_3$  are set to 1.0, 0.5, 0.2 respectively according to empirical study. Besides, the batch size,  $m^s$ , and  $m^t$  are set to 128, 0.2, and 0.2, respectively. We train our whole MMCDA model for 100 epochs with batch size of 16 and early stopping strategy. Parameter optimization is performed by Adam optimizer with learning rate  $4 \times 10^{-4}$  for ActivityNet Captions and Charades-STA and  $3 \times 10^{-4}$  for TACoS, and linear decay of learning rate and gradient clipping of 1.0.

##### C. Baseline Comparison

To evaluate the effectiveness of our proposed MMCDA, we compare it to several state-of-the-art VMR models, which are divided into two categories by the viewpoints of fully-supervised (FS) and weakly-supervised (WS) models. (i) FS models: 2D-TAN [88], ABLR [89], ACRN [90], ACL-K [35], CBP [91], CMIN [42], CSMGAN [37], CTRL [84], GDP [92], SAP [34], SCDM [59], SM-RL [93], MCN [94], TGN [58], VSLNet [87]. (ii) WS models: CTF [45], LoGAN [44], RTBPN [43], SCN [49], TGA [47], MARN [50], VGN [12], VLANet [48].

Following [3], [56], we directly cite the results of compared VMR methods from corresponding works, which are trained and tested on the same datasets in FS or WS settings. Similarly, "CD" means "cross-domain". Note that no weakly-supervised method reports its results on TACoS. The best results are **bold**.

##### D. Does Our Domain Adaption Strategy Work?

To fairly compare with these single-domain baseline models, we serve our proposed MMCDA as a plug-and-play module for several state-of-the-art models (2D-TAN [88], CSMGAN [37], and CBLN [28]) for investigating its effectiveness. Specifically, we develop the following three settings in Table II:

TABLE II

OUR PROPOSED MMCDASERVES AS A PLUG-AND-PLAY MODULE FOR SEVERAL STATE-OF-THE-ART MODELS ON DIFFERENT TRANSFER TASKS.

Model	Setting	T→A				A→C				C→T			
		R@1	R@1	R@5	R@5	R@1	R@1	R@5	R@5	R@1	R@1	R@5	R@5
		IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5
2D-TAN [88]	BL	37.29	25.32	57.81	45.04	59.45	44.51	77.13	61.96	53.25	39.81	85.15	79.33
	SO	32.68	20.10	52.46	42.12	53.51	40.72	72.03	55.12	40.76	36.34	78.73	73.16
	<b>Ours</b>	<b>41.47</b>	<b>28.83</b>	<b>60.37</b>	<b>49.53</b>	<b>71.92</b>	<b>52.98</b>	<b>98.76</b>	<b>89.52</b>	<b>61.02</b>	<b>46.21</b>	<b>90.47</b>	<b>84.33</b>
CSMGAN [37]	BL	33.90	27.09	53.98	41.22	68.52	49.11	87.68	77.43	78.43	60.04	95.43	89.01
	SO	30.07	21.90	50.12	37.81	61.97	43.62	81.95	70.18	73.74	55.93	87.82	79.72
	<b>Ours</b>	<b>39.41</b>	<b>33.65</b>	<b>63.18</b>	<b>50.69</b>	<b>72.84</b>	<b>53.86</b>	<b>99.02</b>	<b>89.96</b>	<b>82.26</b>	<b>66.04</b>	<b>97.28</b>	<b>92.16</b>
CBLN [28]	BL	38.98	27.65	59.96	46.24	66.34	48.12	88.91	79.32	62.43	47.94	93.56	88.20
	SO	33.06	24.39	53.82	42.63	61.43	43.67	80.72	71.84	55.68	42.39	85.32	80.16
	<b>Ours</b>	<b>43.72</b>	<b>35.74</b>	<b>65.71</b>	<b>49.68</b>	<b>73.66</b>	<b>55.60</b>	<b>99.57</b>	<b>89.91</b>	<b>71.03</b>	<b>51.32</b>	<b>96.08</b>	<b>90.73</b>
Model	Setting	C→A				A→T				T→C			
		R@1	R@1	R@5	R@5	R@1	R@1	R@5	R@5	R@1	R@1	R@5	R@5
		IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5
2D-TAN [88]	BL	53.25	39.81	85.15	79.33	59.45	44.51	77.13	61.96	37.29	25.32	57.81	45.04
	SO	50.27	34.82	76.75	74.18	48.26	37.52	68.31	49.35	30.49	18.76	50.38	40.77
	<b>Ours</b>	<b>79.25</b>	<b>62.13</b>	<b>95.83</b>	<b>89.17</b>	<b>63.48</b>	<b>49.76</b>	<b>87.21</b>	<b>78.32</b>	<b>48.74</b>	<b>35.82</b>	<b>86.31</b>	<b>70.06</b>
CSMGAN [37]	BL	78.43	60.04	95.43	89.01	68.52	49.11	87.68	77.43	33.90	27.09	53.98	41.22
	SO	72.92	53.48	86.74	78.94	60.83	44.06	82.34	71.03	31.52	20.71	51.34	38.03
	<b>Ours</b>	<b>84.38</b>	<b>63.99</b>	<b>97.83</b>	<b>93.72</b>	<b>70.51</b>	<b>53.80</b>	<b>79.23</b>	<b>80.16</b>	<b>49.14</b>	<b>36.52</b>	<b>65.86</b>	<b>71.38</b>
CBLN [28]	BL	62.43	47.94	93.56	88.20	66.34	48.12	88.91	79.32	38.98	27.65	59.96	46.24
	SO	54.96	41.72	84.05	80.32	62.07	44.01	79.82	72.05	32.94	23.97	54.06	43.01
	<b>Ours</b>	<b>85.13</b>	<b>64.75</b>	<b>98.26</b>	<b>94.98</b>	<b>72.98</b>	<b>52.87</b>	<b>89.94</b>	<b>80.96</b>	<b>49.25</b>	<b>38.29</b>	<b>87.32</b>	<b>71.54</b>

1) **Baseline (BL)**: we directly report the performance of these baseline models. This is under a single-domain setting, where their training and testing are on the same source dataset.

2) **Source Only (SO)**: we only use the annotated source dataset to train a baseline model, and directly test the trained model on another unannotated target dataset. This is under a simple cross-domain setting.

3) **Ours**: we add our multi-modal cross-domain alignment loss ( $\gamma_1 \mathcal{L}_{DA} + \gamma_2 \mathcal{L}_{MA} + \gamma_3 \mathcal{L}_{SA}$ ) to 2D-TAN, CSMGAN, and CBLN, which can make them adaptive to our transfer tasks. This is under our cross-domain setting.

From Table II, we conclude the following observations:

- Directly testing the trained SO model on the target dataset will severely degenerate the retrieval performance, since there are a large domain discrepancy between two datasets and a huge semantic gap between videos and queries in the target dataset.
- Our domain adaption strategy is effective for these single-domain models, outperforming these baselines and their SO variants by a large margin.

It shows that our multi-modal cross-domain alignment can bring large improvement on single-domain methods.

### E. Performance Comparison and Analysis

**Comparison on TACoS.** As shown in Table III, we compare our cross-domain MMCDAS with the state-of-the-art fully-supervised and weakly-supervised methods on the TACoS dataset, where MMCDAS reaches the highest results in terms of all metrics. Particularly, compared with the best fully-supervised model CSMGAN, MMCDAS(A→T) achieves 12.64% and 10.30% absolute improvements in the strict metrics “R@1, IoU=0.1” and “R@5, IoU=0.5”, respectively; MMCDAS(C→T) obtains 5.23% and 1.40% absolute improvements in “R@1, IoU=0.1” and “R@5, IoU=0.5”, respectively.

The main reason is that in this dataset, the cooking activities appear in the same kitchen scene with some slightly varied cooking objects. Thus, these baseline models are difficult to localize such fine-grained activities correctly. Different from these single-domain models, MMCDAS(A→T) and MMCDAS(C→T) can learn these fine-grained activity representations from ActivityNet Captions and Charades-STA datasets, respectively. Besides, MMCDAS(A→T) and MMCDAS(C→T) also outperform MMCDAS(T→T) with a clear margin, which shows the effectiveness of our multi-modal cross-domain alignment module in the VMR task.

**Comparison on Charades-STA.** We also compare our cross-domain MMCDAS(A→C) with fully-supervised and weakly-supervised models on the Charades-STA dataset in Table IV, where MMCDAS achieves a new state-of-the-art performance on all R@1 and R@5 metrics. Specifically, compared with the best fully-supervised method VSLNet, our MMCDAS(A→C) brings 7.28% and 5.58% improvements in the strict metrics “R@1, IoU=0.3” and “R@1, IoU=0.7”, respectively. Compared with the best weakly-supervised method LoGAN, our MMCDAS(A→C) brings 21.23% and 21.91% improvements in “R@1, IoU=0.7” and “R@5, IoU=0.7”, respectively. It verifies the benefits of transferring the annotation knowledge from the ActivityNet Captions dataset to the Charades-STA dataset through our multi-modal cross-domain alignment module.

**Comparison on ActivityNet Captions.** Table V also reports the grounding performance on the ActivityNet Captions dataset. Obviously, our MMCDAS(C→A) also performs better than supervised methods in all the cases with significant improvements. Particularly, compared with the fully-supervised approach SCDM, our MMCDAS(C→A) obtains 26.33% and 18.64% improvements in the strict metrics “R@1, IoU=0.7” and “R@5, IoU=0.7”, respectively. Compared with the weakly-supervised approach SCN, our MMCDAS(C→A) brings 23.46% and 23.71% improvements in “R@1, IoU=0.5”



TABLE III  
PERFORMANCE ON THE TACoS DATASET.

Model	Mode	R@1 IoU=0.1	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.1	R@5 IoU=0.3	R@5 IoU=0.5
2D-TAN [88]	FS	47.59	37.29	25.32	70.31	57.81	45.04
ABLR [89]	FS	34.70	19.50	9.40	-	-	-
ACRN [90]	FS	24.22	19.52	14.62	47.42	34.97	24.88
CBP [91]	FS	-	27.31	24.79	-	43.64	37.40
CMIN [42]	FS	32.48	24.64	18.05	62.13	38.46	27.02
CSMGAN [37]	FS	42.74	33.90	27.09	68.97	53.98	41.22
GDP [92]	FS	39.68	24.14	13.50	-	-	-
SCDM [59]	FS	-	26.11	21.17	-	40.16	32.18
TGN [58]	FS	41.87	21.77	18.90	53.40	39.06	31.02
VSLNet [87]	FS	-	29.61	24.27	-	-	-
MMCDA(T→T)	FS	36.58	22.63	14.95	54.01	45.77	34.78
MMCDA(A→T)	CD	<b>55.38</b>	<b>43.07</b>	<b>31.69</b>	<b>77.54</b>	<b>63.81</b>	<b>51.52</b>
MMCDA(C→T)	CD	47.97	35.41	26.88	69.53	58.10	42.62

TABLE IV  
PERFORMANCE ON THE CHARADES-STA DATASET.

Model	Mode	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.3	R@5 IoU=0.5	R@5 IoU=0.7
2D-TAN [88]	FS	-	39.81	23.25	-	79.33	52.15
ACL-K [35]	FS	-	30.48	12.20	-	64.84	35.13
SAP [34]	FS	-	27.42	13.36	-	66.37	38.15
SM-RL [93]	FS	-	24.36	11.17	-	61.25	32.08
VSLNet [87]	FS	64.30	47.31	30.19	-	-	-
LoGAN [44]	WS	51.67	34.68	14.54	92.74	74.30	39.11
RTBPN [43]	WS	60.04	32.36	13.24	97.48	71.85	41.18
SCN [49]	WS	42.96	23.58	9.97	95.56	71.80	38.87
TGA [47]	WS	29.68	17.04	6.93	83.87	58.17	26.80
VLANet [48]	WS	45.24	31.83	14.17	95.70	82.85	33.09
MMCDA(C→C)	FS	39.42	24.53	10.04	75.21	53.52	30.87
MMCDA(A→C)	CD	<b>71.58</b>	<b>54.80</b>	<b>35.77</b>	<b>98.36</b>	<b>88.90</b>	<b>61.02</b>
MMCDA(T→C)	CD	48.69	35.17	17.28	85.40	69.93	39.74

and “R@5, IoU=0.5”, respectively.

**Analysis.** From the above results, we can observe that: (i) In all the datasets, our proposed MMCDA on these cross-domain settings (A→T, C→T, C→A, T→A, A→C, T→C) performs better than these fully-supervised settings (T→T, A→A, C→C). There are two reasons as follows: Firstly, our proposed MMCDA on the fully-supervised setting is a simple model (only with  $\mathcal{L}_{SL}$ ), which is only used as a baseline approach. Secondly, we focus on the cross-domain setting in this paper. The better performance on our cross-domain setting shows the effectiveness of our MMCDA network. (ii) Although our MMCDA is under the cross-domain setting, it can significantly outperforms fully-supervised and weakly-supervised VMR methods. It is because MMCDA can effectively transfer the annotation knowledge from the source domain to the target domain. (iii) For different cross-domain VMR tasks (same target dataset and different source datasets), MMCDA performs differently. For example, MMCDA performs better in the A→C task than in the T→C task. One possible reason is that the Activity Captions dataset contains multiple and sufficient scenarios that are able to generalize MMCDA to other datasets. In the contrast, the TACoS dataset only contains cooking videos (monotonous and simple) with a small number of videos, which may limit the generalization ability of our MMCDA.

TABLE V  
PERFORMANCE ON THE ACTIVITYNET CAPTIONS DATASET.

Model	Mode	R@1 IoU=0.3	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.3	R@5 IoU=0.5	R@5 IoU=0.7
ACRN [90]	FS	49.70	31.67	11.25	76.50	60.34	38.57
CTRL [84]	FS	47.43	29.01	10.34	75.32	59.17	37.54
MCN [94]	FS	39.35	21.36	6.43	68.12	53.23	29.70
SCDM [59]	FS	54.80	36.75	19.86	77.29	64.99	41.53
TGN [58]	FS	45.51	28.47	-	57.32	43.33	-
CTF [45]	WS	44.30	23.60	-	-	-	-
RTBPN [43]	WS	49.77	29.63	-	79.89	60.56	-
SCN [49]	WS	47.23	29.22	-	71.45	55.69	-
VGN [12]	WS	46.17	28.79	-	71.23	55.13	-
MARN [50]	WS	47.01	29.95	-	72.02	57.49	-
MMCDA(A→A)	FS	40.18	16.72	11.73	43.81	36.74	30.52
MMCDA(C→A)	CD	<b>70.25</b>	<b>52.68</b>	<b>46.19</b>	<b>88.73</b>	<b>79.40</b>	<b>60.17</b>
MMCDA(T→A)	CD	42.58	20.90	14.69	58.75	46.81	40.13

### F. Ablation Study

In this section, we will perform in-depth ablation studies to evaluate the effectiveness of each component in our MMCDA on different transfer tasks.

**Main ablation study.** To systematically evaluate the effectiveness of each module of our proposed MMCDA, we first conduct an main ablation study on all six transfer tasks. Table VI shows the corresponding results. In the main ablation study, we design the following ablation models: 1) **MMCDA(w/o<sup>3</sup> DA)**: we remove the domain alignment module (Eq. (4)) from MMCDA; 2) **MMCDA(w/o MA)**: we remove the cross-modal alignment module (Eq. (8)) from MMCDA; 3) **MMCDA(w/o SA)**: we remove the specific alignment module (Eq. (11)) from MMCDA; 4) **MMCDA(full)**: our full MMCDA model.

From the Table VI, we can observe that MMCDA(full) outperforms other ablation models in all the tasks, which shows the effectiveness of each module. Especially, compared with MMCDA(w/o DA), MMCDA(full) brings 9.11% improvement in “R@5, IoU=0.5” on the A→T transfer task; compared with MMCDA(w/o MA), MMCDA(full) raises the performance around 7.51% in “R@5, IoU=0.3”; compared with MMCDA(w/o SA), MMCDA(full) obtains the performance improvement about 5.41% in “R@5, IoU=0.1”. It is because the domain alignment module is able to close the domain discrepancy between different datasets for transferring the annotation knowledge; the cross-modal alignment module can reduce the semantic gaps between videos and queries in each dataset for matching a given query and the corresponding video; the specific alignment module aims to learn more fine-grained and representative frame-wise features for more accurate localization.

Based on the above results, we can obtain the following conclusions: 1) Compared with each model, the full model perform the best on all the datasets, which illustrates the effectiveness of each alignment module of our MMCDA. 2) Almost all ablation models outperform all state-of-the-art methods, which shows that our MMCDA can effectively handle the cross-domain video moment retrieval task. Meanwhile, it also illustrates that the excellent performance of our MMCDA does not rely on one specific key alignment module, and our full

<sup>3</sup>In our experiments, “w/o” means “without”, while “w/” means “with”.

TABLE VI  
MAIN ABLATION STUDY ON ALL THE TASKS.

Model	A→T				T→C				C→A			
	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
MMCDA(w/o DA)	34.62	23.50	54.79	42.41	39.84	23.76	76.98	58.73	61.96	43.87	75.42	67.49
MMCDA(w/o MA)	36.61	25.64	56.30	44.19	43.95	23.89	82.91	65.77	64.26	46.82	79.51	70.86
MMCDA(w/o SA)	38.57	27.98	58.45	46.66	47.06	30.84	84.05	66.15	68.72	50.78	83.45	77.46
<b>MMCDA(full)</b>	<b>43.07</b>	<b>31.69</b>	<b>63.81</b>	<b>51.52</b>	<b>48.69</b>	<b>35.17</b>	<b>85.40</b>	<b>69.93</b>	<b>70.25</b>	<b>52.68</b>	<b>88.73</b>	<b>79.40</b>

Model	T→A				C→T				A→C			
	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5
MMCDA(w/o DA)	38.54	16.27	52.85	40.06	31.70	20.46	51.06	36.79	67.15	46.91	94.02	82.38
MMCDA(w/o MA)	39.04	17.68	54.72	41.65	33.46	22.73	54.93	37.01	68.42	47.88	96.35	84.72
MMCDA(w/o SA)	40.35	19.79	57.21	44.92	34.87	23.96	56.49	39.20	69.19	49.96	97.83	86.54
<b>MMCDA(full)</b>	<b>42.58</b>	<b>20.90</b>	<b>58.75</b>	<b>46.81</b>	<b>35.41</b>	<b>26.88</b>	<b>58.10</b>	<b>42.62</b>	<b>71.58</b>	<b>54.80</b>	<b>98.36</b>	<b>88.90</b>

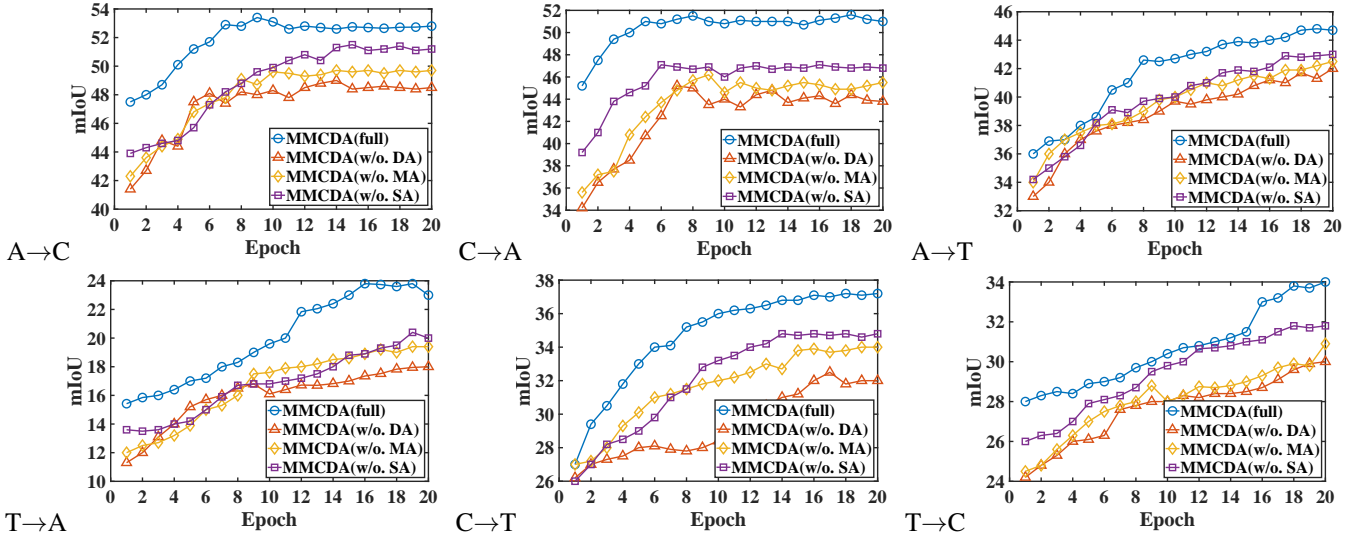


Fig. 4. Performance of each ablation module on the target domain during training epochs.

model is robust to address this cross-domain task. 3) We can observe that both MMCDA(w/o SA) and MMCDA(w/o MA) outperforms MMCDA(w/o DA). It is because the domain alignment is the key of the cross-domain VMR task to transfer the knowledge from source domain to target domain.

**Training process of different ablation models.** Following [49], we conduct six experiments to analyze the training process of different ablation models on each transfer task. Fig. 4 shows the results, where “mIoU” is the evaluation metric. We can obtain the following representative observations: (i) On each epoch, MMCDA(full) outperforms ablation models, which further demonstrates the effectiveness of each module. (ii) MMCDA(full) performs better in the A→C task than in the C→A task. It is because ActivityNet Captions (20,000 videos) contains more annotated videos (annotation knowledge) than Charades-STA (9,848 videos). (iii) For each transfer task, MMCDA(full) converges faster than ablation models, which shows that the full model is more efficient on time-consuming.

**Analysis on domain alignment and cross-modal alignment.** As shown in Table VII, we also conduct ablation study within the domain alignment and cross-modal alignment examine the effectiveness of each loss function. For the domain alignment, the video-based domain alignment loss  $\mathcal{L}_{DV}$  and the query-

based domain alignment loss  $\mathcal{L}_{DQ}$  bring significant improvement in all metrics on both tasks. Especially,  $\mathcal{L}_{DV}$  brings 8.36% improvement in “R@1, IoU=0.5” on the A→C transfer task, and raises the performance around 7.97% in “R@5, IoU=0.3” on the C→A transfer task. Besides,  $\mathcal{L}_{DQ}$  obtains the performance improvement about 6.43% in “R@5, IoU=0.5” on the A→C transfer task, and brings 7.40% improvement in “R@5, IoU=0.3” on the C→A transfer task. As for the cross-modal alignment, the cross-modal consistent loss  $\mathcal{L}_{M1}$  and the cross-modal feature distribution loss  $\mathcal{L}_{M2}$  also effectively improve the grounding accuracy. For example, on the A→C transfer task,  $\mathcal{L}_{M1}$  raises the performance around 4.24% and 4.84% in “R@1, IoU=0.3” and “R@5, IoU=0.5” respectively, while it achieves the performance improvement about 4.32% and 6.10% in “R@1, IoU=0.5” and “R@5, IoU=0.3” respectively on the C→A transfer task. Moreover,  $\mathcal{L}_{M2}$  brings 2.42% improvement in “R@1, IoU=0.5” on the A→C transfer task, and brings 5.62% improvement in “R@5, IoU=0.3” on the C→A transfer task.

**Analysis on hyperparameters.** As shown in Table VIII, we analyze the impact of three hyperparameters:  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$ . It can be observed that, with the increase of  $\gamma_1$ ,  $\gamma_2$  and  $\gamma_3$ , their performance follows a general trend, i.e., rises at

TABLE VII  
ABLATION STUDY ON THE DOMAIN ALIGNMENT AND CROSS-MODAL ALIGNMENT ON ALL THE TASKS.

Module	Changes	A→C						C→A					
		R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7
Domain Alignment	w/ $\mathcal{L}_{DV}$	<b>71.58</b>	<b>54.80</b>	<b>35.77</b>	<b>98.36</b>	<b>88.90</b>	<b>61.02</b>	<b>70.25</b>	<b>52.68</b>	<b>46.19</b>	<b>88.73</b>	<b>79.40</b>	<b>60.17</b>
	w/o $\mathcal{L}_{DV}$	65.83	44.32	30.89	90.15	81.30	53.37	64.24	44.99	39.51	80.76	72.09	53.83
	w/ $\mathcal{L}_{DQ}$	<b>71.58</b>	<b>54.80</b>	<b>35.77</b>	<b>98.36</b>	<b>88.90</b>	<b>61.02</b>	<b>70.25</b>	<b>52.68</b>	<b>46.19</b>	<b>88.73</b>	<b>79.40</b>	<b>60.17</b>
Cross-modal Alignment	w/o $\mathcal{L}_{DQ}$	66.34	47.38	30.85	92.31	82.47	54.94	65.03	45.38	40.12	81.33	73.24	54.63
	w/ $\mathcal{L}_{M1}$	<b>71.58</b>	<b>54.80</b>	<b>35.77</b>	<b>98.36</b>	<b>88.90</b>	<b>61.02</b>	<b>70.25</b>	<b>52.68</b>	<b>46.19</b>	<b>88.73</b>	<b>79.40</b>	<b>60.17</b>
	w/o $\mathcal{L}_{M1}$	68.34	48.96	32.75	95.17	84.06	59.08	67.62	48.36	43.75	82.63	75.90	55.41
	w/ $\mathcal{L}_{M2}$	<b>71.58</b>	<b>54.80</b>	<b>35.77</b>	<b>98.36</b>	<b>88.90</b>	<b>61.02</b>	<b>70.25</b>	<b>52.68</b>	<b>46.19</b>	<b>88.73</b>	<b>79.40</b>	<b>60.17</b>
	w/o $\mathcal{L}_{M2}$	70.03	50.21	33.84	97.18	86.35	60.38	68.42	49.17	44.63	83.01	76.34	57.29

Module	Changes	A→T						T→A					
		R@1, IoU=0.1	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.1	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7
Domain Alignment	w/ $\mathcal{L}_{DV}$	<b>55.38</b>	<b>43.07</b>	<b>31.69</b>	<b>77.54</b>	<b>63.81</b>	<b>51.52</b>	<b>42.58</b>	<b>20.90</b>	<b>14.69</b>	<b>58.75</b>	<b>46.81</b>	<b>40.13</b>
	w/o $\mathcal{L}_{DV}$	53.42	40.81	27.35	73.80	60.35	48.26	39.97	18.96	13.21	54.96	43.28	37.95
	w/ $\mathcal{L}_{DQ}$	<b>55.38</b>	<b>43.07</b>	<b>31.69</b>	<b>77.54</b>	<b>63.81</b>	<b>51.52</b>	<b>42.58</b>	<b>20.90</b>	<b>14.69</b>	<b>58.75</b>	<b>46.81</b>	<b>40.13</b>
Cross-modal Alignment	w/o $\mathcal{L}_{DQ}$	53.94	41.12	28.75	74.28	61.37	48.93	40.26	19.35	13.62	55.78	44.89	38.74
	w/ $\mathcal{L}_{M1}$	<b>55.38</b>	<b>43.07</b>	<b>31.69</b>	<b>77.54</b>	<b>63.81</b>	<b>51.52</b>	<b>42.58</b>	<b>20.90</b>	<b>14.69</b>	<b>58.75</b>	<b>46.81</b>	<b>40.13</b>
	w/o $\mathcal{L}_{M1}$	54.26	41.38	28.94	75.62	61.53	49.04	40.72	19.46	13.95	56.27	45.08	38.96
	w/ $\mathcal{L}_{M2}$	<b>55.38</b>	<b>43.07</b>	<b>31.69</b>	<b>77.54</b>	<b>63.81</b>	<b>51.52</b>	<b>42.58</b>	<b>20.90</b>	<b>14.69</b>	<b>58.75</b>	<b>46.81</b>	<b>40.13</b>
	w/o $\mathcal{L}_{M2}$	54.03	41.27	29.05	75.73	61.45	49.72	40.93	19.32	13.87	56.94	45.36	39.18

Module	Changes	C→T						T→C					
		R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.1	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7
Domain Alignment	w/ $\mathcal{L}_{DV}$	<b>47.97</b>	<b>35.41</b>	<b>26.88</b>	<b>69.53</b>	<b>58.10</b>	<b>42.62</b>	<b>48.69</b>	<b>35.17</b>	<b>17.28</b>	<b>85.40</b>	<b>69.93</b>	<b>39.74</b>
	w/o $\mathcal{L}_{DV}$	45.74	33.02	23.43	66.08	55.49	39.90	45.03	33.72	16.35	83.92	67.30	35.98
	w/ $\mathcal{L}_{DQ}$	<b>47.97</b>	<b>35.41</b>	<b>26.88</b>	<b>69.53</b>	<b>58.10</b>	<b>42.62</b>	<b>48.69</b>	<b>35.17</b>	<b>17.28</b>	<b>85.40</b>	<b>69.93</b>	<b>39.74</b>
Cross-modal Alignment	w/o $\mathcal{L}_{DQ}$	46.02	32.84	23.72	65.11	55.76	40.31	45.32	33.60	16.03	83.75	67.84	36.20
	w/ $\mathcal{L}_{M1}$	<b>47.97</b>	<b>35.41</b>	<b>26.88</b>	<b>69.53</b>	<b>58.10</b>	<b>42.62</b>	<b>48.69</b>	<b>35.17</b>	<b>17.28</b>	<b>85.40</b>	<b>69.93</b>	<b>39.74</b>
	w/o $\mathcal{L}_{M1}$	46.84	33.78	24.13	66.49	56.17	40.96	45.43	33.85	16.84	84.12	68.01	36.87
	w/ $\mathcal{L}_{M2}$	<b>47.97</b>	<b>35.41</b>	<b>26.88</b>	<b>69.53</b>	<b>58.10</b>	<b>42.62</b>	<b>48.69</b>	<b>35.17</b>	<b>17.28</b>	<b>85.40</b>	<b>69.93</b>	<b>39.74</b>
	w/o $\mathcal{L}_{M2}$	47.16	34.02	24.75	66.14	56.85	41.77	46.37	34.12	16.99	84.24	68.72	37.41

TABLE VIII  
ABLATION STUDY ON HYPERPARAMETERS  $\gamma_1$ ,  $\gamma_2$ ,  $\gamma_3$ .

$\gamma_1$	A→T				T→C				C→A			
	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
0.8	42.89	30.04	61.31	50.92	47.62	33.09	83.01	68.26	68.83	50.12	87.90	77.32
0.9	43.01	31.45	62.14	51.03	48.17	34.92	84.89	68.91	69.47	51.83	88.20	78.74
<b>1.0</b>	<b>43.07</b>	<b>31.69</b>	<b>63.81</b>	<b>51.52</b>	<b>48.69</b>	<b>35.17</b>	<b>85.40</b>	<b>69.93</b>	<b>70.25</b>	<b>52.68</b>	<b>88.73</b>	<b>79.40</b>
1.1	42.31	30.07	62.19	50.10	47.98	34.03	84.92	68.33	69.15	51.32	87.39	78.10
1.2	40.24	28.19	60.32	49.88	46.63	33.77	83.45	67.51	68.06	50.87	86.58	77.93

$\gamma_2$	A→T				T→C				C→A			
	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
0.3	41.10	30.25	62.11	50.19	46.07	33.91	83.15	67.83	68.49	50.14	86.46	77.03
0.4	42.11	31.12	62.98	51.17	47.86	34.40	84.52	69.02	69.19	51.30	87.98	78.13
<b>0.5</b>	<b>43.07</b>	<b>31.69</b>	<b>63.81</b>	<b>51.52</b>	<b>48.69</b>	<b>35.17</b>	<b>85.40</b>	<b>69.93</b>	<b>70.25</b>	<b>52.68</b>	<b>88.73</b>	<b>79.40</b>
0.6	41.39	30.18	62.90	50.30	47.26	33.32	84.41	68.89	69.97	51.28	87.26	78.74
0.7	40.92	29.83	61.47	50.01	46.74	32.19	83.38	68.42	69.02	50.10	86.04	77.31

$\gamma_3$	A→T				T→C				C→A			
	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5	R@1 IoU=0.3	R@1 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.5
0.1	42.33	31.02	63.15	50.90	47.96	34.03	83.47	68.99	69.72	51.03	87.52	78.49
<b>0.2</b>	<b>43.07</b>	<b>31.69</b>	<b>63.81</b>	<b>51.52</b>	<b>48.69</b>	<b>35.17</b>	<b>85.40</b>	<b>69.93</b>	<b>70.25</b>	<b>52.68</b>	<b>88.73</b>	<b>79.40</b>
0.3	42.25	30.91	62.90	50.93	48.07	34.16	84.91	69.18	68.37	51.48	87.96	79.01
0.4	41.93	29.85	62.04	50.17	47.82	34.03	84.25	68.72	68.09	51.19	87.14	78.17
0.5	41.31	29.11	61.69	49.82	46.79	33.25	83.62	67.04	66.27	50.53	85.52	76.38

first and then starts to decline. It is because when these three hyperparameters are relatively small, our MMCDA pays more attention to the fully-supervised learning during the pre-

training phase in the source domain, and it can obtain more annotation knowledge from the source domain. As the increase of these hyperparameters, MMCDA adds the attention to the

TABLE IX  
ABLATION STUDY ON THE VIDEO AND QUERY ENCODERS FOR ALL THE TASKS.

Module	Changes	A→C						C→A					
		R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7
Video Encoder	w/ Bi-GRU	<b>71.58</b>	<b>54.80</b>	<b>35.77</b>	<b>98.36</b>	<b>88.90</b>	<b>61.02</b>	<b>70.25</b>	<b>52.68</b>	<b>46.19</b>	<b>88.73</b>	<b>79.40</b>	<b>60.17</b>
	w/o Bi-GRU	69.83	53.01	33.87	96.02	87.42	60.58	68.94	50.36	43.51	87.25	78.02	59.13
	w/ Self-attention	<b>71.58</b>	<b>54.80</b>	<b>35.77</b>	<b>98.36</b>	<b>88.90</b>	<b>61.02</b>	<b>70.25</b>	<b>52.68</b>	<b>46.19</b>	<b>88.73</b>	<b>79.40</b>	<b>60.17</b>
	w/o Self-attention	69.42	53.76	34.89	97.08	86.31	59.64	68.02	51.13	44.26	86.98	77.86	58.40
Query Encoder	w/ Bi-GRU	<b>71.58</b>	<b>54.80</b>	<b>35.77</b>	<b>98.36</b>	<b>88.90</b>	<b>61.02</b>	<b>70.25</b>	<b>52.68</b>	<b>46.19</b>	<b>88.73</b>	<b>79.40</b>	<b>60.17</b>
	w/o Bi-GRU	68.60	51.72	34.17	96.50	87.16	59.93	68.36	51.47	44.82	87.05	78.14	59.37
	w/ Self-attention	<b>71.58</b>	<b>54.80</b>	<b>35.77</b>	<b>98.36</b>	<b>88.90</b>	<b>61.02</b>	<b>70.25</b>	<b>52.68</b>	<b>46.19</b>	<b>88.73</b>	<b>79.40</b>	<b>60.17</b>
	w/o Self-attention	69.28	52.25	33.75	96.72	85.92	59.13	68.74	51.26	45.30	86.41	77.59	58.81
Module	Changes	A→T						T→A					
		R@1, IoU=0.1	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.1	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7
Video Encoder	w/ Bi-GRU	<b>55.38</b>	<b>43.07</b>	<b>31.69</b>	<b>77.54</b>	<b>63.81</b>	<b>51.52</b>	<b>42.58</b>	<b>20.90</b>	<b>14.69</b>	<b>58.75</b>	<b>46.81</b>	<b>40.13</b>
	w/o Bi-GRU	54.35	42.61	30.04	76.53	62.90	50.43	40.99	19.02	13.18	56.39	44.72	38.26
	w/ Self-attention	<b>55.38</b>	<b>43.07</b>	<b>31.69</b>	<b>77.54</b>	<b>63.81</b>	<b>51.52</b>	<b>42.58</b>	<b>20.90</b>	<b>14.69</b>	<b>58.75</b>	<b>46.81</b>	<b>40.13</b>
	w/o Self-attention	53.97	42.14	29.87	75.32	61.37	49.92	40.26	19.23	13.58	56.30	45.12	39.28
Query Encoder	w/ Bi-GRU	<b>55.38</b>	<b>43.07</b>	<b>31.69</b>	<b>77.54</b>	<b>63.81</b>	<b>51.52</b>	<b>42.58</b>	<b>20.90</b>	<b>14.69</b>	<b>58.75</b>	<b>46.81</b>	<b>40.13</b>
	w/o Bi-GRU	54.03	41.78	29.05	77.06	61.85	49.26	39.15	18.46	12.87	56.04	44.83	38.71
	w/ Self-attention	<b>55.38</b>	<b>43.07</b>	<b>31.69</b>	<b>77.54</b>	<b>63.81</b>	<b>51.52</b>	<b>42.58</b>	<b>20.90</b>	<b>14.69</b>	<b>58.75</b>	<b>46.81</b>	<b>40.13</b>
	w/o Self-attention	54.40	41.92	29.30	75.48	61.02	49.85	39.55	18.94	13.70	56.81	44.92	38.45
Module	Changes	C→T						T→C					
		R@1, IoU=0.1	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.1	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.3	R@5, IoU=0.5	R@5, IoU=0.7
Video Encoder	w/ Bi-GRU	<b>47.97</b>	<b>35.41</b>	<b>26.88</b>	<b>69.53</b>	<b>58.10</b>	<b>42.62</b>	<b>48.69</b>	<b>35.17</b>	<b>17.28</b>	<b>85.40</b>	<b>69.93</b>	<b>39.74</b>
	w/o Bi-GRU	45.23	34.50	24.77	67.35	56.84	41.05	46.82	34.06	15.27	83.82	67.19	37.58
	w/ Self-attention	<b>47.97</b>	<b>35.41</b>	<b>26.88</b>	<b>69.53</b>	<b>58.10</b>	<b>42.62</b>	<b>48.69</b>	<b>35.17</b>	<b>17.28</b>	<b>85.40</b>	<b>69.93</b>	<b>39.74</b>
	w/o Self-attention	45.83	34.26	25.19	68.03	57.15	41.33	47.26	33.08	16.42	84.96	68.49	37.80
Query Encoder	w/ Bi-GRU	<b>47.97</b>	<b>35.41</b>	<b>26.88</b>	<b>69.53</b>	<b>58.10</b>	<b>42.62</b>	<b>48.69</b>	<b>35.17</b>	<b>17.28</b>	<b>85.40</b>	<b>69.93</b>	<b>39.74</b>
	w/o Bi-GRU	46.08	34.12	23.75	67.82	56.71	40.79	46.34	34.18	16.85	84.11	67.82	37.43
	w/ Self-attention	<b>47.97</b>	<b>35.41</b>	<b>26.88</b>	<b>69.53</b>	<b>58.10</b>	<b>42.62</b>	<b>48.69</b>	<b>35.17</b>	<b>17.28</b>	<b>85.40</b>	<b>69.93</b>	<b>39.74</b>
	w/o Self-attention	45.16	33.71	24.06	67.44	56.23	40.10	47.02	33.28	15.35	83.74	68.42	36.85

TABLE X  
ABLATION STUDY ON THE INTRA-SAMPLE DISTRIBUTION FUNCTION  $\mu(\cdot)$  AND THE INTER-SAMPLE DISTRIBUTION FUNCTION  $\sigma(\cdot)$  ON ALL THE TASKS.

Loss	Changes	A→C		C→A		A→T		T→A		C→T		T→C	
		R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5	R@1, IoU=0.3	R@1, IoU=0.5
$\mathcal{L}_{DV}$	w/ $\mu(\cdot)$	<b>71.58</b>	<b>54.80</b>	<b>70.25</b>	<b>52.68</b>	<b>43.07</b>	<b>31.69</b>	<b>42.58</b>	<b>20.90</b>	<b>35.41</b>	<b>26.88</b>	<b>48.69</b>	<b>35.17</b>
	w/o $\mu(\cdot)$	69.31	51.27	68.97	51.35	41.88	28.46	39.20	18.72	33.98	25.03	46.85	32.34
	w/ $\sigma(\cdot)$	<b>71.58</b>	<b>54.80</b>	<b>70.25</b>	<b>52.68</b>	<b>43.07</b>	<b>31.69</b>	<b>42.58</b>	<b>20.90</b>	<b>35.41</b>	<b>26.88</b>	<b>48.69</b>	<b>35.17</b>
	w/o $\sigma(\cdot)$	69.16	51.59	68.84	51.42	41.76	27.42	38.99	19.03	34.05	24.87	46.62	32.01
$\mathcal{L}_{DQ}$	w/ $\mu(\cdot)$	<b>71.58</b>	<b>54.80</b>	<b>70.25</b>	<b>52.68</b>	<b>43.07</b>	<b>31.69</b>	<b>42.58</b>	<b>20.90</b>	<b>35.41</b>	<b>26.88</b>	<b>48.69</b>	<b>35.17</b>
	w/o $\mu(\cdot)$	69.84	51.02	69.24	51.76	42.03	29.70	39.58	19.17	34.25	24.99	47.06	32.94
	w/ $\sigma(\cdot)$	<b>71.58</b>	<b>54.80</b>	<b>70.25</b>	<b>52.68</b>	<b>43.07</b>	<b>31.69</b>	<b>42.58</b>	<b>20.90</b>	<b>35.41</b>	<b>26.88</b>	<b>48.69</b>	<b>35.17</b>
	w/o $\sigma(\cdot)$	68.92	50.40	68.84	51.03	41.92	28.85	39.01	19.22	34.07	25.12	47.23	33.85
$\mathcal{L}_{M1}$	w/ $\mu(\cdot)$	<b>71.58</b>	<b>54.80</b>	<b>70.25</b>	<b>52.68</b>	<b>43.07</b>	<b>31.69</b>	<b>42.58</b>	<b>20.90</b>	<b>35.41</b>	<b>26.88</b>	<b>48.69</b>	<b>35.17</b>
	w/o $\mu(\cdot)$	70.04	49.76	67.81	50.35	42.01	29.53	39.60	19.28	34.59	25.05	47.35	33.29
	w/ $\sigma(\cdot)$	<b>71.58</b>	<b>54.80</b>	<b>70.25</b>	<b>52.68</b>	<b>43.07</b>	<b>31.69</b>	<b>42.58</b>	<b>20.90</b>	<b>35.41</b>	<b>26.88</b>	<b>48.69</b>	<b>35.17</b>
	w/o $\sigma(\cdot)$	69.42	49.25	68.34	51.05	42.17	30.88	40.56	19.04	34.72	25.06	47.38	34.01

multi-modal cross-domain alignment module during the main training phase, which helps MMCDa transfer the annotation knowledge from the source domain to the target domain. However, when these hyperparameters are too large, MMCDa may have difficulty in obtaining the annotation knowledge, which will lead the performance drop. Therefore, in our paper, we set  $\gamma_1 = 1.0$ ,  $\gamma_2 = 0.5$  and  $\gamma_3 = 0.2$ , where our MMCDa can obtain the best performance.

**Analysis on the multi-modal encoders.** As shown in Table IX, we investigate different variants of multi-modal encoders on all the tasks. In the A→C task, for the video encoder, we can observe that Bi-GRU brings the improvement of 1.19%, 1.76%, 1.90%, 2.34%, 1.48% and 0.44% in all metrics, which

demonstrates that Bi-GRU can extract the sequential information among these consecutive frames. Also, the self-attention module improves performance (2.16%, 1.04%, 0.88%, 1.28%, 2.59% and 1.38%) to the full model. The performance boost is due to its ability to capture the intra-video contexts. As for the query encoder, we could find that if we remove the Bi-GRU module, the full model will decrease the performance of 2.30%, 2.08%, 1.60%, 1.64%, 1.74% and 1.09% in all metrics, which presents the effectiveness of Bi-GRU in incorporating the contextual textual information. Moreover, the self-attention module improves the accuracy of 2.30%, 2.51%, 2.02%, 1.64%, 2.98% and 1.89% since it can obtain the semantic dependencies in the long query.

TABLE XI

PERFORMANCE COMPARISON OF OUR MMCDA WITH STATE-OF-THE-ART MULTI-MODAL CROSS-DOMAIN ALGORITHMS ON ALL THE TASKS. NOTE THAT WE ONLY COMPARE OUR MMCDA TO STATE-OF-THE-ART OPEN-SOURCE ALGORITHMS. IN [95], [96], THESE AUTHORS DO NOT NAME THEIR MODELS, FOR CONVENIENCE, WE NAME THEIR DESIGNED “CROSS-DOMAIN FAKE NEWS DETECTION USING MULTI-MODAL DATA” FRAMEWORK AS “CFND” AND “CROSS-MODAL RETRIEVAL AND MODEL ADAPTATION” AS “CRMA”.

Model	A→C		C→A		A→T		T→A		C→T		T→C	
	R@1,	R@1,	R@1,	R@1,	R@1,	R@1,	R@1,	R@1,	R@1,	R@1,	R@1,	R@1,
	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5	IoU=0.3	IoU=0.5
xMUDA [97]	66.33	46.94	66.10	48.02	38.45	26.92	36.45	17.26	30.17	23.10	40.52	29.33
MM-SADA [56]	67.56	48.05	67.30	48.86	37.97	26.43	36.08	18.25	31.50	23.38	44.81	30.15
PMC [98]	68.91	48.36	68.27	49.87	39.72	26.05	38.24	17.83	30.81	23.96	45.86	29.74
CrossCLR [99]	69.07	48.41	68.54	50.08	40.13	26.84	38.50	18.15	32.96	24.11	45.25	30.82
CFND [95]	69.12	49.50	68.71	50.23	41.06	26.45	37.82	18.01	33.42	24.53	45.84	31.53
CRMA [96]	69.35	50.88	68.84	50.59	40.85	27.06	38.94	18.23	33.86	24.72	46.50	31.95
<b>Our MMCDA</b>	<b>71.58</b>	<b>54.80</b>	<b>70.25</b>	<b>52.68</b>	<b>43.07</b>	<b>31.69</b>	<b>42.58</b>	<b>20.90</b>	<b>35.41</b>	<b>26.88</b>	<b>48.69</b>	<b>35.17</b>

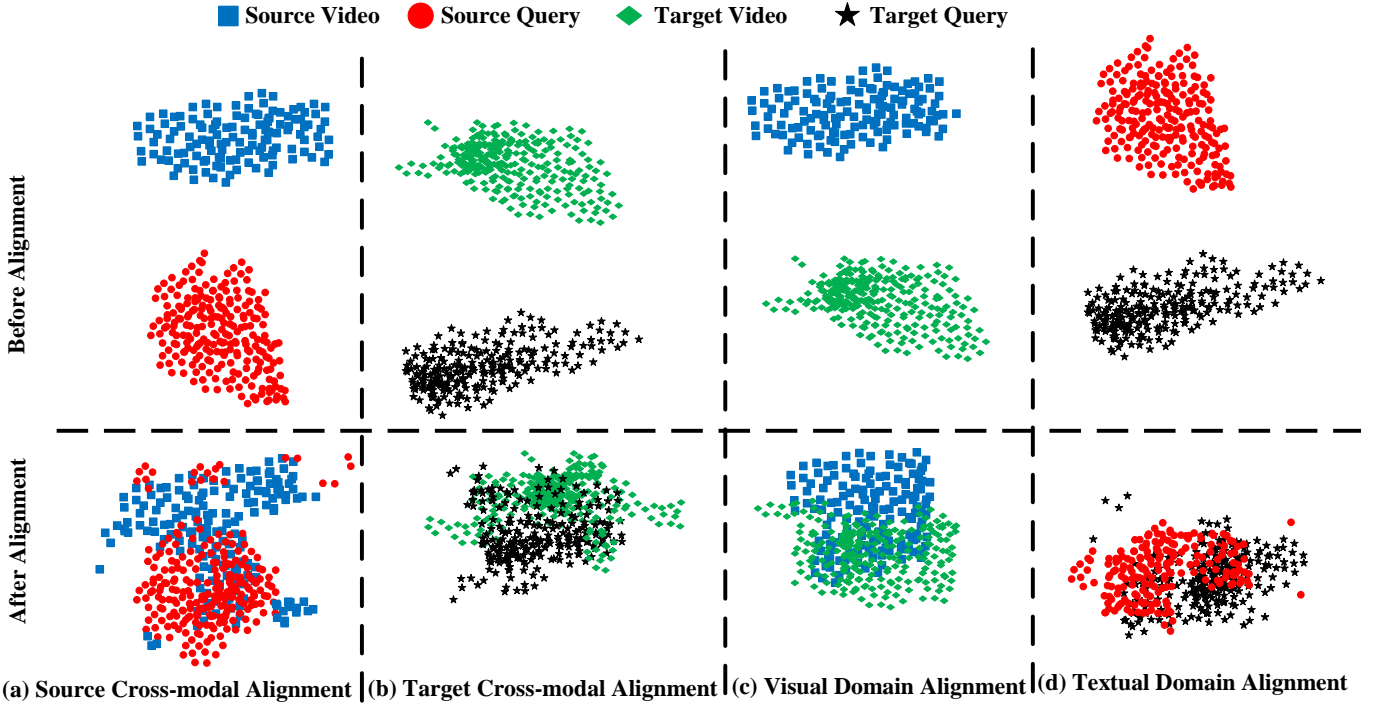


Fig. 5. T-SNE visualization on cross-modal cross-domain features on the A→C task. (a) and (b) are the visualization of cross-modal alignment for individual domains, and each domain contains the features from videos and queries. Similarly, in (c) and (d), for each modality, we visualize the domain alignment, where the features from both source and target domains are utilized. In (e), we integrate all the features from two domains and two modalities by our cross-modal cross-domain alignment framework. Best viewed in color and each color means a kind of feature.

#### Analysis on intra-and inter-sample distribution functions.

To evaluate the power of our newly proposed distribution functions (intra-sample function  $\mu(\cdot)$  and inter-sample function  $\sigma(\cdot)$ ), we conduct an ablation study on all the tasks. Since only three losses ( $\mathcal{L}_{DV}$ ,  $\mathcal{L}_{DQ}$  and  $\mathcal{L}_{M1}$ ) contains  $\mu(\cdot)$  and  $\sigma(\cdot)$ , we analyze the performance of  $\mu(\cdot)$  and  $\sigma(\cdot)$  on these three losses. As shown in Table X, both  $\mu(\cdot)$  and  $\sigma(\cdot)$  improve the performance by large margins. Particularly, on the A→C task,  $\mu(\cdot)$  and  $\sigma(\cdot)$  of  $\mathcal{L}_{DV}$  achieve the performance improvement by 2.27% and 2.42% in terms of “R@1, IoU=0.3” respectively. As for the C→A task,  $\mu(\cdot)$  and  $\sigma(\cdot)$  of  $\mathcal{L}_{M1}$  improve the retrieval performance by 2.39% and 2.03% “R@1, IoU=0.7” respectively. The significant performance shows the effectiveness of  $\mu(\cdot)$  and  $\sigma(\cdot)$  on the domain alignment and cross-modal alignment.

#### G. Do Other Multi-modal Cross-domain methods Work Well?

Someone might ask if other multi-modal cross-domain methods can also work well. To evaluate the performance of different multi-modal cross-domain methods in the VMR task, we compare our MMCDA with multiple multi-modal cross-domain methods (xMUDA [97], MM-SADA [56], and PMC [98]) on different tasks. Table XI reports experimental results.

Since these state-of-the-art multi-modal cross-domain methods cannot be directly used for the VMR task, we first add our pre-training loss  $\mathcal{L}_{SL}$  to learn the annotation knowledge. Then, we use these methods to find some possible moment-query pairs in the target domain. Finally, we utilize our inference to obtain the target moment boundary of these methods.

From Table XI, for the A→C and C→A tasks, we can notice that our MMCDA outperforms other methods by a large margin. Especially, compared with the best baseline



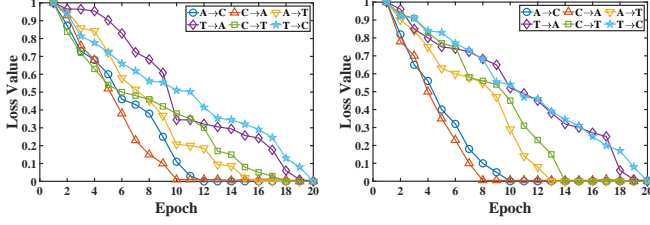


Fig. 6. Visualization of domain gap on different transfer tasks for each modality during training epochs (left: visual domain gap, right: textual domain gap). Note that we use regularized loss function values as the metric of the domain gap, where larger loss value means the larger domain gap. “loss value=1” means the largest domain gap and “loss value=0” means the smallest domain gap. Best viewed in color.

method CRMA, MMCDA raises the performance around 2.23%, 3.92%, 1.41% and 2.09% in all metrics. The significant improvement further shows the effectiveness of MMCDA.

#### H. Visualization on Cross-modal Cross-domain Alignment

Since cross-modal alignment is an important module, we provide the cross-modal attention map to analyze the semantic alignment between visual and textual representations. As shown in Fig. 5, based on our proposed MMCDA, video can match its corresponding query with high similarity.

In Fig. 6, the faster convergence curve indicates that our method is better able to reduce the domain gap of the two datasets in the transfer task. For example, in the A→C and C→A tasks, their corresponding curves converge within 12 epochs (for the visual domain gap) and 10 epochs (for the textual domain gap). It is because our MMCDA can effectively and efficiently transfer the annotation knowledge from the source dataset to the target dataset by our domain alignment module, which closes the domain gap between the ActivityNet Captions and Charades-STA datasets. When we use the TACoS dataset as the source dataset, the curves in the T→C and T→A tasks often converge slowly. The main reason is that the TACoS dataset contains significantly little annotation knowledge, which limits the performance on the target datasets. On the contrary, when we utilize the TACoS dataset as the target dataset, our MMCDA can obtain enough annotation knowledge by the domain alignment, which speeds the curve convergence.

#### I. Qualitative Analysis

To qualitatively investigate the effectiveness of our MMCDA, we report some representative examples from three datasets on six transfer tasks as shown in Fig. 7-9, where the ground truth and the fully-supervised method 2D-TAN are two baselines. Obviously, our MMCDA localizes more accurately than 2D-TAN, which verifies the effectiveness of MMCDA.

For our designed ablation models, we can find the following interesting points: (i) MMCDA(full) achieves more precise localization than each ablation model, which shows the effectiveness of each alignment module. (ii) Compared with MMCDA(w/o. DA), MMCDA(full) achieves significant improvement because MMCDA(full) leverages the domain

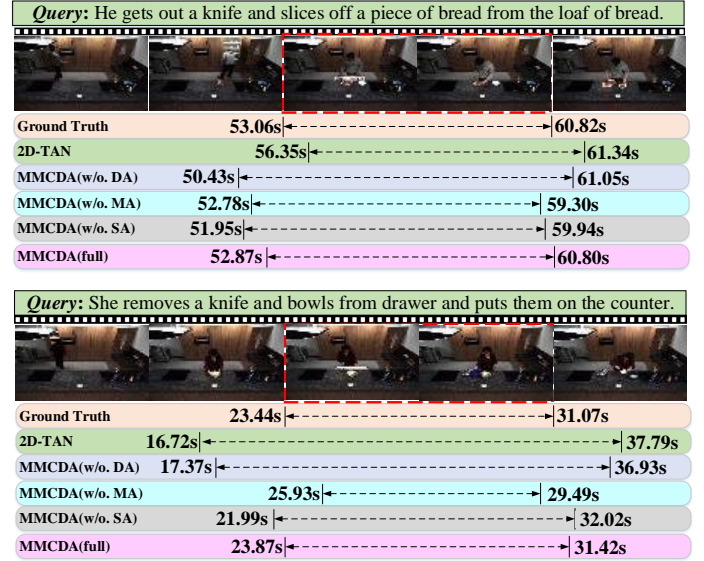


Fig. 7. Qualitative results sampled from the TACoS dataset (top: A→T, bottom: C→T).




Fig. 8. Qualitative results sampled from the Charades-STA dataset (top: A→C, bottom: T→C).

alignment to alleviate the domain discrepancy between source and target domains. (iii) Compared with MMCDA(w/o. MA), MMCDA(full) obtains satisfactory improvement, since MMCDA(full) utilizes the cross-modal alignment to bridge the semantic gaps between videos and queries in the target domain. (iv) Compared with MMCDA(w/o. SA), MMCDA(full) also gains better performance. One possible reason is that MMCDA(full) employs the specific alignment to obtain a more accurate moment boundary.

#### V. CONCLUSION

In this paper, we make the first attempt for the cross-domain video moment retrieval task. Thanks to our well-designed three alignment modules that align data distribution, video-query features, and context semantics across the



<b>Query:</b> The man cleans again with soap and water the car for the second time.		
		
Ground Truth	31.16s	64.91s
2D-TAN	26.35s	67.85s
MMCDA(w/o. DA)	25.87s	66.94s
MMCDA(w/o. MA)	32.05s	65.43s
MMCDA(w/o. SA)	31.95s	65.98s
MMCDA(full)	31.07s	64.80s

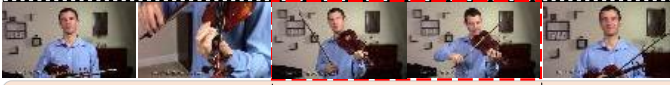
<b>Query:</b> He continues playing the instrument while pausing to speak to the camera.		
		
Ground Truth	138.65s	215.94s
2D-TAN	150.72s	201.39s
MMCDA(w/o. DA)	113.26s	213.02s
MMCDA(w/o. MA)	149.33s	218.37s
MMCDA(w/o. SA)	131.93s	223.46s
MMCDA(full)	137.58s	214.83s

Fig. 9. Qualitative results sampled from the ActivityNet Captions dataset (top: C→A, bottom: T→A).

source and target domains, our proposed MMCDA can easily adapt to unseen domain data without manual annotations. Extensive experiments demonstrate that MMCDA despite being cross-domain significantly outperforms state-of-the-art single-domain methods. In the future, we will apply MMCDA to other tasks/datasets [29], [100] to further improve its generalization. Meanwhile, exploring different ways to align source/target domains to improve the model is also our aimed work.

## REFERENCES

- [1] G. Wang, X. Xu, F. Shen, H. Lu, Y. Ji, and H. T. Shen, "Cross-modal dynamic networks for video moment retrieval with text query," *IEEE TMM*, 2022.
- [2] Y. Wang, M. Liu, Y. Wei, Z. Cheng, Y. Wang, and L. Nie, "Siamese alignment network for weakly supervised video moment retrieval," *IEEE TMM*, 2022.
- [3] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, and R. S. M. Goh, "Natural language video localization: A revisit in span-based question answering framework," *IEEE TPAMI*, 2021.
- [4] H. Tang, J. Zhu, M. Liu, Z. Gao, and Z. Cheng, "Frame-wise cross-modal matching for video moment retrieval," *IEEE TMM*, 2021.
- [5] J. Gao and C. Xu, "Learning video moment retrieval without a single annotated video," *TCSVT*, 2021.
- [6] X. Yang, S. Wang, J. Dong, J. Dong, M. Wang, and T.-S. Chua, "Video moment retrieval with cross-modal neural architecture search," *IEEE TIP*, 2022.
- [7] J. Teng, X. Lu, Y. Gong, X. Liu, X. Nie, and Y. Yin, "Regularized two granularity loss function for weakly supervised video moment retrieval," *IEEE TMM*, 2021.
- [8] Z. Zhang, Z. Zhao, Z. Zhang, Z. Lin, Q. Wang, and R. Hong, "Temporal textual localization in video via adversarial bi-directional interaction networks," *IEEE TMM*, vol. 23, pp. 3306–3317, 2020.
- [9] Y. Zeng, D. Cao, S. Lu, H. Zhang, J. Xu, and Z. Qin, "Moment is important: Language-based video moment retrieval via adversarial learning," *ACM TOMM*, vol. 18, no. 2, pp. 1–21, 2022.
- [10] F. Zhu, Y. Zhu, X. Chang, and X. Liang, "Vision-language navigation with self-supervised auxiliary reasoning tasks," in *CVPR*, 2020, pp. 10012–10022.
- [11] J. Singha, A. Roy, and R. H. Laskar, "Dynamic hand gesture recognition using vision-based approach for human-computer interaction," *Neural Computing and Applications*, vol. 29, no. 4, pp. 1129–1141, 2018.
- [12] Z. Zhang, Z. Zhao, Z. Lin, X. He *et al.*, "Counterfactual contrastive learning for weakly-supervised vision-language grounding," *NeurIPS*, 2020.
- [13] M. Soldan, M. Xu, S. Qu, J. Tegner, and B. Ghanem, "Vlg-net: Video-language graph matching network for video grounding," in *ICCV*, 2021, pp. 3224–3234.
- [14] M. Zhang, Y. Yang, X. Chen, Y. Ji, X. Xu, J. Li, and H. T. Shen, "Multi-stage aggregated transformer network for temporal language localization in videos," in *CVPR*, 2021, pp. 12 669–12 678.
- [15] Z. Guo, Z. Zhao, W. Jin, W. Dazhou, L. Ruitao, and J. Yu, "Tao-highlight: Commodity-aware multi-modal video highlight detection in e-commerce," *IEEE TMM*, 2021.
- [16] X. Fang, D. Liu, P. Zhou, Z. Xu, and R. Li, "Hierarchical local-global transformer for temporal sentence grounding," *arXiv preprint arXiv:2208.14882*, 2022.
- [17] R. Zeng, C. Gan, P. Chen, W. Huang, Q. Wu, and M. Tan, "Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization," *IEEE TIP*, pp. 5797–5808, 2019.
- [18] J. Lin, L.-Y. Duan, S. Wang, Y. Bai, Y. Lou, V. Chandrasekhar, T. Huang, A. Kot, and W. Gao, "Hnlp: Compact deep invariant representations for video matching, localization, and retrieval," *IEEE TMM*, vol. 19, no. 9, pp. 1968–1983, 2017.
- [19] C. G. Snoek, B. Huurnink, L. Hollink, M. De Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE TMM*, vol. 9, no. 5, pp. 975–986, 2007.
- [20] X. Fang, Y. Hu, P. Zhou, and D. O. Wu, "Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat," *IEEE TETCI*, 2021.
- [21] S. C. Hoi and M. R. Lyu, "A multimodal and multilevel ranking scheme for large-scale video retrieval," *IEEE TMM*, vol. 10, no. 4, pp. 607–619, 2008.
- [22] X. Fang, Y. Hu, P. Zhou, and D. O. Wu, "Animc: A soft approach for auto-weighted noisy and incomplete multi-view clustering," *IEEE TAI*, 2021.
- [23] X. Fang, Y. Hu, P. Zhou, and D. O. Wu, "V3h: View variation and view heredity for incomplete multiview clustering," *IEEE TAI*, vol. 1, no. 3, pp. 233–247, 2020.
- [24] Y. Wang, J. Deng, W. Zhou, and H. Li, "Weakly supervised temporal adjacent network for language grounding," *IEEE TMM*, 2021.
- [25] Y. Zeng, D. Cao, X. Wei, M. Liu, Z. Zhao, and Z. Qin, "Multi-modal relational graph for cross-modal video moment retrieval," in *CVPR*, 2021.
- [26] D. Liu, X. Qu, J. Dong, and P. Zhou, "Adaptive proposal generation network for temporal sentence localization in videos," in *EMNLP*, 2021.
- [27] D. Cao, Y. Zeng, M. Liu, X. He, M. Wang, and Z. Qin, "Strong: Spatio-temporal reinforcement learning for cross-modal video moment localization," in *MM*, 2020, pp. 4162–4170.
- [28] D. Liu, X. Qu, J. Dong, P. Zhou, Y. Cheng, W. Wei, Z. Xu, and Y. Xie, "Context-aware biaffine localizing network for temporal sentence grounding," in *CVPR*, June 2021, pp. 11 235–11 244.
- [29] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvr: A large-scale dataset for video-subtitle moment retrieval," in *ECCV*, 2020.
- [30] Y. Zhao, Z. Zhao, Z. Zhang, and Z. Lin, "Cascaded prediction network via segment tree for temporal video grounding," in *CVPR*, 2021, pp. 4197–4206.
- [31] S. Xiao, L. Chen, S. Zhang, W. Ji, J. Shao, L. Ye, and J. Xiao, "Boundary proposal network for two-stage natural language video localization," in *AAAI*, 2021.
- [32] H. Wang, Z.-J. Zha, L. Li, D. Liu, and J. Luo, "Structured multi-level interaction network for video moment localization via language query," in *CVPR*, 2021.
- [33] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *AAAI*, 2019, pp. 9062–9069.
- [34] S. Chen and Y. Jiang, "Semantic proposal for activity localization in videos via sentence query," in *AAAI*, 2019, pp. 8199–8206.
- [35] R. Ge, J. Gao, K. Chen, and R. Nevatia, "Mac: Mining activity concepts for language-based temporal localization," in *WACV*, 2019, pp. 245–253.
- [36] D. Liu, X. Qu, J. Dong, and P. Zhou, "Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network," in *COLING*, 2020, pp. 1841–1851.
- [37] D. Liu, X. Qu, X. Liu, J. Dong, P. Zhou, and Z. Xu, "Jointly cross- and self-modal graph attention network for query-based moment localization," in *MM*, 2020, pp. 4070–4078.

- [38] K. Li, D. Guo, and M. Wang, "Proposal-free video grounding with contextual pyramid network," in *AAAI*, 2021, pp. 1902–1910.
- [39] C. Rodriguez, E. Marrese-Taylor, F. S. Saleh, H. Li, and S. Gould, "Proposal-free temporal moment localization of a natural-language query in video using guided attention," in *WACV*, 2020.
- [40] J. Wang, L. Ma, and W. Jiang, "Temporally grounding language queries in videos by contextual boundary-aware prediction," in *AAAI*, 2020.
- [41] D. Zhang, X. Dai, X. Wang, Y. Wang, and L. S. Davis, "MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment," in *CVPR*, 2019, pp. 1247–1257.
- [42] Z. Zhang, Z. Lin, Z. Zhao, and Z. Xiao, "Cross-modal interaction networks for query-based moment retrieval in videos," in *SIGIR*, 2019, pp. 655–664.
- [43] Z. Zhang, Z. Lin, Z. Zhao, J. Zhu, and X. He, "Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos," in *MM*, 2020, pp. 4098–4106.
- [44] R. Tan, H. Xu, K. Saenko, and B. A. Plummer, "Logan: Latent graph co-attention network for weakly-supervised video moment retrieval," in *WACV*, 2021.
- [45] Z. Chen, L. Ma, W. Luo, and K. K. Wong, "Weakly-supervised spatio-temporally grounding natural sentence in video," in *ACL*, 2019.
- [46] X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, and J. Huang, "Weakly supervised dense event captioning in videos," in *NeurIPS*, 2018.
- [47] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury, "Weakly supervised video moment retrieval from text queries," in *CVPR*, 2019.
- [48] M. Ma, S. Yoon, J. Kim, Y. Lee, S. Kang, and C. D. Yoo, "Vlanet: Video-language alignment network for weakly-supervised video moment retrieval," in *ECCV*, 2020, pp. 156–171.
- [49] Z. Lin, Z. Zhao, Z. Zhang, Q. Wang, and H. Liu, "Weakly-supervised video moment retrieval via semantic completion network," in *AAAI*, 2020.
- [50] Y. Song, J. Wang, L. Ma, Z. Yu, and J. Yu, "Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos," *arXiv preprint arXiv:2003.07048*, 2020.
- [51] P. Wang, Y. Yang, Y. Xia, K. Wang, X. Zhang, and S. Wang, "Information maximizing adaptation network with label distribution priors for unsupervised domain adaptation," *IEEE TMM*, 2022.
- [52] A. Zhang, Y. Yang, J. Xu, X. Cao, X. Zhen, and L. Shao, "Latent domain generation for unsupervised domain adaptation object counting," *IEEE TMM*, 2022.
- [53] M. Jing, L. Meng, J. Li, L. Zhu, and H. T. Shen, "Adversarial mixup ratio confusion for unsupervised domain adaptation," *IEEE TMM*, 2022.
- [54] P. Wang, C. Ding, W. Tan, M. Gong, K. Jia, and D. Tao, "Uncertainty-aware clustering for unsupervised domain adaptive object re-identification," *IEEE TMM*, 2022.
- [55] Y. Tao, J. Zhang, J. Hong, and Y. Zhu, "Dreamt: Diversity enlarged mutual teaching for unsupervised domain adaptive person re-identification," *IEEE TMM*, 2022.
- [56] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *CVPR*, 2020, pp. 122–132.
- [57] Y. Wang, S. Qiu, D. Li, C. Du, B.-L. Lu, and H. He, "Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition," *JAS*, 2022.
- [58] K. Chen, R. Kovvuri, and R. Nevatia, "Query-guided regression network with context policy for phrase grounding," in *ICCV*, 2017, pp. 824–832.
- [59] Y. Yuan, L. Ma, J. Wang, W. Liu, and W. Zhu, "Semantic conditioned dynamic modulation for temporal sentence grounding in videos," in *NeurIPS*, 2019.
- [60] J. Na, H. Jung, H. J. Chang, and W. Hwang, "Fixbi: Bridging domain spaces for unsupervised domain adaptation," in *CVPR*, 2021.
- [61] Y. Jing, W. Wang, L. Wang, and T. Tan, "Cross-modal cross-domain moment alignment network for person search," in *CVPR*, 2020, pp. 10675–10683.
- [62] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *CVPR*, 2019.
- [63] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [64] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "Augmented cyclic adversarial learning for low resource domain adaptation," in *ICLR*, 2018.
- [65] X. Huang, Y. Peng, and M. Yuan, "Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE TCYB*, pp. 1047–1059, 2018.
- [66] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *NIPS*, 2006.
- [67] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel two-sample test," *JMLR*, vol. 13, pp. 723–773, 2012.
- [68] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *NIPS*, 2016, pp. 343–351.
- [69] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017, pp. 2962–2971.
- [70] X. Song, S. Zhao, J. Yang, H. Yue, P. Xu, R. Hu, and H. Chai, "Spatio-temporal contrastive domain adaptation for action recognition," in *CVPR*, 2021, pp. 9787–9795.
- [71] D. Kim, Y.-H. Tsai, B. Zhuang, X. Yu, S. Sclaroff, K. Saenko, and M. Chandraker, "Learning cross-modal contrastive features for video domain adaptation," in *ICCV*, 2021, pp. 13618–13627.
- [72] Q. Chen, Y. Liu, and S. Albanie, "Mind-the-gap! unsupervised domain adaptation for text-video retrieval," in *AAAI*, 2021, pp. 1072–1080.
- [73] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.
- [74] F. Long, T. Yao, Q. Dai, X. Tian, J. Luo, and T. Mei, "Deep domain adaptation hashing with adversarial learning," in *ACM SIGIR*, 2018, pp. 725–734.
- [75] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *NeurIPS*, vol. 31, 2018.
- [76] X. Li, Y. He, F. Fioranelli, and X. Jing, "Semisupervised human activity recognition with radar micro-doppler signatures," *IEEE TGRS*, 2021.
- [77] D. Guan, J. Huang, A. Xiao, S. Lu, and Y. Cao, "Uncertainty-aware unsupervised domain adaptation in object detection," *IEEE TMM*, 2021.
- [78] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.
- [79] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
- [80] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS*, 2014.
- [81] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.
- [82] Z. Wang, L. He, X. Tu, J. Zhao, X. Gao, S. Shen, and J. Feng, "Robust video-based person re-identification by hierarchical mining," *IEEE TCSVT*, 2021.
- [83] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *TACL*, vol. 1, pp. 25–36, 2013.
- [84] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: temporal activity localization via language query," in *ICCV*, 2017, pp. 5277–5285.
- [85] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *ICCV*, 2017, pp. 706–715.
- [86] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *ECCV*, 2016, pp. 510–526.
- [87] H. Zhang, A. Sun, W. Jing, and J. T. Zhou, "Span-based localizing network for natural language video localization," in *ACL*, 2020, pp. 6543–6554.
- [88] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2d temporal adjacent networks for moment localization with natural language," in *AAAI*, 2020.
- [89] Y. Yuan, T. Mei, and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *AAAI*, 2019.
- [90] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T. Chua, "Attentive moment retrieval in videos," in *SIGIR*, 2018, pp. 15–24.
- [91] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *CVPR*, 2018, pp. 1307–1315.
- [92] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP*, 2014, pp. 1724–1734.
- [93] W. Wang, Y. Huang, and L. Wang, "Language-driven temporal activity localization: A semantic matching reinforcement learning model," in *CVPR*, 2019.
- [94] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with temporal language," in *EMNLP*, 2018.

- [95] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, “Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data,” in *AAAI*, vol. 35, no. 1, 2021, pp. 557–565.
- [96] W. Zhao, X. Wu, and J. Luo, “Cross-domain image captioning via cross-modal retrieval and model adaptation,” *IEEE TIP*, vol. 30, pp. 1180–1192, 2021.
- [97] M. Jaritz, T.-H. Vu, R. d. Charette, E. Wirbel, and P. Pérez, “xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation,” in *CVPR*, 2020, pp. 12 605–12 614.
- [98] W. Zhang, D. Xu, J. Zhang, and W. Ouyang, “Progressive modality cooperation for multi-modality domain adaptation,” *IEEE TIP*, vol. 30, pp. 3293–3306, 2021.
- [99] M. Zolfaghari, Y. Zhu, P. Gehler, and T. Brox, “Crossclr: Cross-modal contrastive learning for multi-modal video representations,” in *ICCV*, 2021, pp. 1450–1459.
- [100] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, “Hero: Hierarchical encoder for video+ language omni-representation pre-training,” in *EMNLP*, 2020.