# Dual Alignment Unsupervised Domain Adaptation for Video-Text Retrieval

Xiaoshuai Hao[1,2], Wanqian Zhang[1]*, Dayan Wu[1], Fei Zhu[1,2], Bo Li[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{haoxiaoshuai,zhangwanqian,wudayan,zhufei,libo}@iie.ac.cn

## Abstract

*Video-text retrieval is an emerging stream in both computer vision and natural language processing communities, which aims to find relevant videos given text queries. In this paper, we study the notoriously challenging task, i.e., Unsupervised Domain Adaptation Video-text Retrieval (UDAVR), wherein training and testing data come from different distributions. Previous works merely alleviate the domain shift, which however overlook the pairwise misalignment issue in target domain, i.e., there exist no semantic relationships between target videos and texts. To tackle this, we propose a novel method named Dual Alignment Domain Adaptation (DADA). Specifically, we first introduce the cross-modal semantic embedding to generate discriminative source features in a joint embedding space. Besides, we utilize the video and text domain adaptations to smoothly balance the minimization of the domain shifts. To tackle the pairwise misalignment in target domain, we propose the Dual Alignment Consistency (DAC) to fully exploit the semantic information of both modalities in target domain. The proposed DAC adaptively aligns the video-text pairs which are more likely to be relevant in target domain, enabling that positive pairs are increasing progressively and the noisy ones will potentially be aligned in the later stages. To that end, our method can generate more truly aligned target pairs and ensure the discriminability of target features. Compared with the state-of-the-art methods, DADA achieves 20.18% and 18.61% relative improvements on R@1 under the setting of TGIF→MSR-VTT and TGIF→MSVD respectively, demonstrating the superiority of our method.*

## 1. Introduction

Video-text retrieval enables users to search videos with a simple and natural language description. The de facto paradigm is to learn high-level visual-textual embeddings by off-the-shelf feature extractors, and to measure semantic similarities in a joint embedding space [13, 42, 46, 63]. Despite their thrilling success, the primary assumption is that training and testing data come from the same distribution, which whereas may not hold in real scenarios.

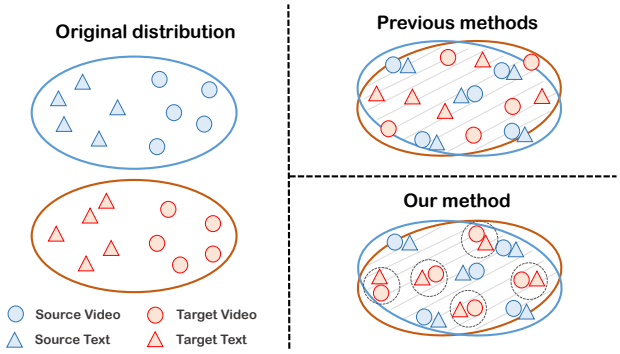



Figure 1. Illustration of the proposed method. Previous methods simply bring source and target features closer (blue and red ovals are overlapping each other), whereas inevitably mixing target videos (red circles) and texts (red triangles) together, ignoring whether they are semantically relevant or not. Instead, our method exploits the semantic structures in target domain to adaptively generate truly aligned video-text pairs (dotted circles) and ensure the discriminability of target data. Best viewed in color.

To alleviate the domain shift problem, Unsupervised Domain Adaptation (UDA) has gained a lot of attention due to its efficient training without the need of supervision in target domain. UDA transfers knowledge from a labeled source domain to an unlabeled target domain [15, 33, 40, 41, 53], which has made remarkable progress in many fields, such as image classification [33, 56], autonomous driving [54, 55], medical image processing [35, 36], and video-based action recognition [50, 52]. However, these methods are originally designed for classification tasks, which might not be suitable for the video-text retrieval.

Note that in UDA Video-text Retrieval (UDAVR), there exists no *identical label set* for source and target domains. The only supervision is the semantic relationship in source dataset, which is also the general setting for UDA

*Corresponding author.

cross-modal tasks [4, 11, 62, 64]. To that end, some approaches have been recently proposed [9, 17, 39], such as directly minimizing the distribution discrepancy [17], distilling knowledge from the source domain [9], or introducing pre-defined prototype assignments [39]. However, they overlook the *pairwise misalignment* issue in target domain, i.e., there exist no semantic relationships between target videos and texts. Merely alleviating the video and text domain shifts is a sub-optimal solution, which fails to fully explore the semantic structures of target data, i.e., whether the video-text pair is semantically relevant or not. As illustrated in Fig. 1, previous methods bring the learned source and target features close together, which whereas inevitably mixes up target videos and texts, ignoring whether they are a truly relevant pair or not. This will further induce less discriminative target features, and thus becomes the motivation of our work.

In this paper, we propose a novel method named Dual Alignment Domain Adaptation (DADA) to tackle the pairwise misalignment issue in target domain. We first introduce the cross-modal semantic embedding to generate discriminative source features in a joint embedding space, where semantically relevant pairs should lie close together and vice versa. To alleviate the domain shift, we further utilize a smooth adaptation procedure to balance the minimization of distribution shifts between source and target domains. Last but not least, to tackle the *pairwise misalignment* in target domain, we propose a simple yet effective Dual Alignment Consistency (DAC), which fully exploits the semantic information of both modalities in target domain. The proposed DAC adaptively aligns the video-text pairs which are more likely to be relevant in target domain, enabling that (1) positive pairs are increasing progressively, (2) the noisy ones will potentially be aligned in the later stages and (3) the discriminability of target features. Extensive experiments on several benchmarks demonstrate the superiority of our method.

The contributions of this paper are mainly threefold:

- To tackle the *pairwise misalignment* problem in UDAVR task, we develop a novel method named Dual Alignment Domain Adaptation (DADA) which fully exploits the semantic structures of target data.
- The proposed Dual Alignment Consistency (DAC) mechanism adaptively aligns the most similar videos and texts in target domain, ensure that the positive pairs are increasing progressively and the noisy ones are potentially aligned in later stages.
- Compared with the state-of-the-art methods, DADA achieves 20.18% and 18.61% relative improvements on R@1 under the setting of TGIF→MSRVTT and TGIF→MSVD respectively, demonstrating the superiority of our method.

## 2. Related Work

**Video-Text Retrieval.** In recent years, cross-modal embedding-based approaches [2, 10, 20, 26, 27, 37, 47, 58] have emerged as a dominant paradigm for video-text retrieval. [48] proposes the JEMC framework using action, object, text and audio features by a simple concatenation fusion strategy. CE [37] adopts video features extracted from all modalities to encode a video. T2VLAD [59] automatically learns text-and-video semantic topics and reemphasizes the importance of local semantic alignment between texts and videos. HGR [10] proposes a Hierarchical Graph Reasoning (HGR) model, which decomposes video-text pairs into global-to-local levels. GPO [5] learns to automatically adapt itself to the best pooling strategy for different baselines. Recently, the Contrastive Language-Image Pretraining (CLIP) [3] model is widely used in video-text retrieval [24, 31, 38, 45]. CLIP4Clip [43] investigates three mechanisms of similarity calculation based on the pre-trained CLIP. Similarly, CLIP2video [18] focuses on the spatial semantics captured by the CLIP model. Different from them, we explore the video-text retrieval task through the lens of unsupervised domain adaptation.

**Unsupervised Domain Adaptation.** UDA transfers predictive models from a fully-labeled source domain to an unlabeled target domain. Existing classification-based UDA methods seek to alleviate the domain shift between source and target domains [15, 22, 33, 40, 41, 56, 60]. Besides, UDA methods have been extended to various video-based tasks, like video action recognition [6, 12, 49], video segmentation [7, 8] and video localisation [1]. Recently, some cross-modal tasks also resort to UDA and try to utilize the unpaired data in target domain, such as image captioning [11, 62, 64] and VQA [4]. The similar work to ours is DCKT [29] which focuses on UDA image-text retrieval and transfers knowledge from a large dataset to promote the model performance on small dataset. However, DCKT needs labeled target image-text pairs during the training procedure, which fails to work well for UDAVR task.

**Unsupervised Domain Adaptation for Video-Text Retrieval.** To the best of our knowledge, there are only a few explorations of the UDAVR task [9, 17, 39]. MAN [17] proposes three alignments to alleviate different gaps in UDAVR task. CAPQ [9] comprises a concept preservation regulariser to enhance the transferability of the learned embeddings. ACP [39] focuses on minimizing both uni-modal and cross-modal distribution shift across the source and target domains. Compared to these methods, our approach differs in three aspects. (1) MAN tries to directly alleviate three different gaps in a classification-based manner, which is not suitable for cross-modal retrieval task. (2) CAPQ and ACP maximize the mutual information or minimize the KL-divergence between the prototype assignments of source and target videos, which however ignores the domain shift

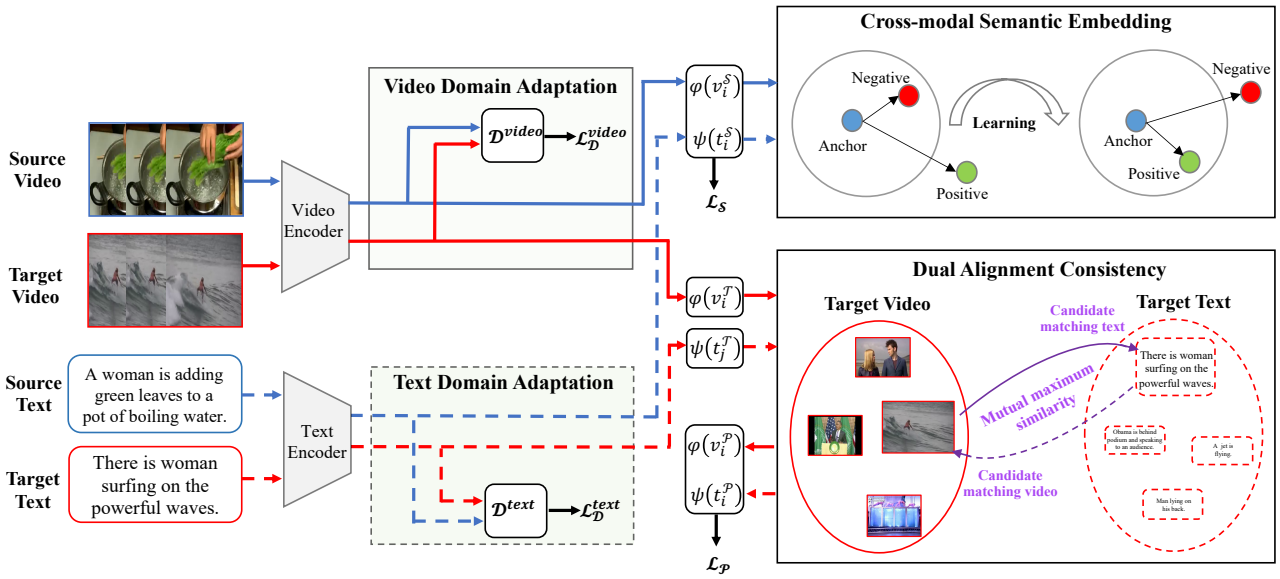



Figure 2. The overall framework of DADA. Video/text features are first fed into video/text encoders to generate high-level representations. The video and text domain adaptation modules simultaneously alleviates the distribution shifts across domains in both modalities ($\mathcal{L}_{\mathcal{D}}^{video}$ and $\mathcal{L}_{\mathcal{D}}^{text}$). Source video and text features are expected to be discriminative by the cross-modal semantic embedding ($\mathcal{L}_{\mathcal{S}}$). Besides, the proposed Dual Alignment Consistency (DAC) adaptively aligns the target video-text pairs which are more likely to be relevant and progressively generates *dual aligned* video-text pairs $(v_i^p, t_i^p)_{i=1}^{n_p}$ in target domain($\mathcal{L}_{\mathcal{P}}$). Best viewed in color.

in text modality. (3) The semantic relationships of videos and texts in target domain have not been fully exploited by previous methods, leading to the pairwise misalignment issue, which is the primary concern of this paper.

## 3. Methodology

### 3.1. Preliminaries

For notational clarity, we first introduce some symbols and definitions used throughout this paper. Formally, assume that we have a set of samples in source domain $\left\{(\mathcal{V}^s, \mathcal{T}^s) = (v_i^s, t_i^s)_{i=1}^{n_s}\right\}$, where $n_s$ indicates the number of video-text pairs. Similarly, we also have a set of samples in target domain $\left\{\mathcal{V}^t = \{v_i^t\}_{i=1}^{n_t}, \mathcal{T}^t = \left\{t_j^t\right\}_{j=1}^{n_t}\right\}$ with two collections of $n_t$ videos $\mathcal{V}^t$ and texts $\mathcal{T}^t$, respectively. Note that the target videos and texts are *unpaired*, which means the supervised information, i.e., *whether one target video-text pair is semantically relevant or not*, is not available in target domain. The Unsupervised Domain Adaptation Video-text Retrieval (UDAVR) aims at improving the model's generalization performance on target domain with the utilization of source domain. The overall framework of our method is illustrated in Fig. 2.

Given one video-text pair, following the state-of-the-art baseline in video-text retrieval [5], we utilize a video encoder $\varphi(\cdot)$ and a text encoder $\psi(\cdot)$ to map each video sample $v$ and text description $t$ into a joint embedding space. The visual embedding $\varphi(v) \in \mathbb{R}^M$ and text embedding $\psi(t) \in \mathbb{R}^M$ are semantically relevant if the text describes the video, where $M$ denotes the dimension in the common space. In the source domain, we utilize the video-text contrastive loss to guide the semantic alignment learning. Following [30, 32, 51], the contrastive loss considers matched pairs as positive and all others pairs that can be formed in a batch as negatives. For each input video-text pair $(v_i, t_i)$, the video-text contrastive loss consists of two symmetric terms, one for video-to-text classification:

$$\mathcal{L}^{v2t} = -log\frac{exp\left(s\left(v_i, t_i\right)/\tau\right)}{\sum_j^B exp\left(s\left(v_i, t_j\right)/\tau\right)}, \tag{1}$$

and the other for text-to-video classification:

$$\mathcal{L}^{t2v} = -log\frac{exp\left(s\left(t_i, v_i\right)/\tau\right)}{\sum_j^B exp\left(s\left(t_i, v_j\right)/\tau\right)}. \tag{2}$$

$\tau$ is the temperature parameter and $B$ is the batch size. We calculate similarity scores with the cosine similarity, which is a widely-used similarity metric and has been proved effective [10, 16]:

$$s\left(v_i, t_j\right) = \frac{\varphi\left(v_i\right) \cdot \psi\left(t_j\right)}{\|\varphi\left(v_i\right)\| \|\psi\left(t_j\right)\|}, \tag{3}$$

where $\varphi\left(v_i\right)$ and $\psi\left(t_j\right)$ are the corresponding mapped features, and $\|\cdot\|$ denotes the $l_2$ norm of vectors and the Frobenius norm of matrices. Formally, the contrastive loss for the video-text pairs is as follows:

$$\mathcal{L}_{\mathcal{S}} = \frac{1}{2}\left(\mathcal{L}^{v2t} + \mathcal{L}^{t2v}\right). \tag{4}$$

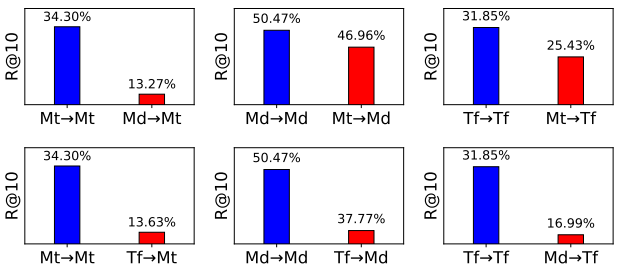



Figure 3. Illustration of the performance decrease when training data and testing data are sampled from different distributions. Mt, Md and Tf denote the dataset MSR-VTT, MSVD and TGIF, respectively. For instance, Mt→Md denotes training on MSR-VTT and testing on MSVD.

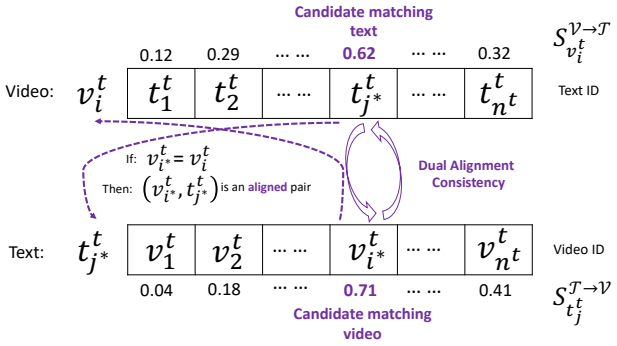



Figure 4. Illustration of Dual Alignment Consistency (DAC). If $v_{i*}^t$ and $t_{j*}^t$ are the reciprocal nearest neighbor of each other, then they can be considered as a truly aligned pair in target domain.

### 3.2. Domain Adaptation in UDAVR

Different datasets usually have inconsistent data distributions and representations, thus leading to the domain shift problem. To verify this, we show the performance comparisons in Fig. 3, where training data and testing data come from different distributions. As can be seen, when both training and testing data come from MSR-VTT dataset, i.e., Mt→Mt, the R@10 result is $34.30\%$. In the contrast, when training on MSVD and testing on MSR-VTT, i.e., Md→Mt, the R@10 result decreases to $13.27\%$, indicating a relative $21.03\%$ performance drop. The significant performance degeneration identifies the domain shift problem in UDAVR.

To alleviate this, we resort to recently proposed DA method [14], which generates intermediate domain representations on-the-fly to gradually bridge the source and target domains. By utilizing the appropriate intermediate domain to bridge the source and target, the source knowledge can be better transferred to the target domain. Specifically, we denote distributions of source, target and intermediate domain as $P_s$, $P_t$ and $P_i$ respectively, and use $D(\cdot)$ to represent the Euclidean distance. Besides, we also introduce the domain factor $\alpha$ for the source and target domains respectively. The domain factor can be seen as the relevance of the intermediate domain to the other two extreme domains. Thus, in the video stream, the distance relationship (contrary to the relevance relationship) between $P_i$ and other two domains, i.e., $P_s$ and $P_t$, can be formulated as:

$$\frac{D(P_s^v, P_i^v)}{D(P_t^v, P_i^v)} = \frac{\alpha}{1-\alpha}. \tag{5}$$

Formally, the domain shift problem can be converted into minimizing the intermediate domain loss as:

$$\mathcal{L}_{\mathcal{D}}^{video} = \alpha D(P_s^v, P_i^v) + (1-\alpha) D(P_t^v, P_i^v). \tag{6}$$

The loss in Eq. 6 aims at guiding the distribution of appropriate intermediate domain to keep the right distance to the source and target domains. Similarly, in the text stream, the intermediate domain loss can be computed as:

$$\mathcal{L}_{\mathcal{D}}^{text} = \alpha D(P_s^t, P_i^t) + (1-\alpha) D(P_t^t, P_i^t). \tag{7}$$

To sum up, the domain adaptation loss can be defined as:

$$\mathcal{L}_{\mathcal{D}} = \mathcal{L}_{\mathcal{D}}^{video} + \mathcal{L}_{\mathcal{D}}^{text}. \tag{8}$$

### 3.3. Dual Alignment Consistency

Despite the efficiency of the cross-modal semantic embedding and domain adaptations, we argue that the desired discriminability in target domain still can not be ensured. Note that in UDA cross-modal tasks, there exists no *identical label set* for source and target domains, and the only supervision available is the semantic relationship in the source dataset [4, 11, 62, 64]. Merely alleviating the domain shift will inevitably mix up target videos and texts, ignoring whether they are a relevant pair or not. This thus leads to the *pairwise misalignment* issue in target domain. In other words, the target videos and texts are *unpaired*, which means the pairwise information is not available in target domain.

To alleviate this, we propose a simple yet effective Dual Alignment Consistency (DAC). The DAC tries to utilize the truly aligned target video-text pairs which are more likely to be semantically relevant, and to avoid including the noisy ones which tend to be irrelevant. Specifically, given the target set $\left\{\mathcal{V}^t = \{v_i^t\}_{i=1}^{n_t}, \mathcal{T}^t = \{t_j^t\}_{j=1}^{n_t}\right\}$, we try to find if there exist truly positive video-text pairs. $(v_i^t, t_j^t)$ can be considered as a truly positive pair if and only if $v_i^t$ and $t_j^t$ are mutually the most similar to each other, indicating a *dual aligned* pair. For a target video $v_i^t$, we calculate the similarities of $v_i^t$ and all the target texts, which can be defined as:

$$\mathbf{S}_{v_i^t}^{\mathcal{V}\to\mathcal{T}} = [S_{v_i^t t_1^t}, S_{v_i^t t_2^t}, ..., S_{v_i^t t_j^t}, ..., S_{v_i^t t_{n^t}^t}]. \tag{9}$$

Table 1. Comparison of three datasets, i.e., MSR-VTT, MSVD and TGIF in the UDAVR task.

| Dataset | #Videos | #Caps. | Video Len. | Query Len. | #Train | #Val | # Test | Scene | Text Src. | Video Src. | Semantics |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MSR-VTT | 10,000 | 200k | 20s | 9.34 | 6,513 | 497 | 2,990 | Open | AMT | YouTube | Category |
| MSVD | 1,970 | 86k | 10s | 7.03 | 1,200 | 100 | 670 | Open | AMT | YouTube | Multi. Lang. |
| TGIF | 101,412 | 120k | 3s | 8.67 | 79,451 | 10,651 | 11,310 | Anim. GIF | Crowdsouring | Tumblr | None |

Then, we nominate the selected text for target video $v_i^t$ with the maximum similarity as the candidate matching text:

$$j^* = \arg \max_{j \in \{1,2,\ldots,n_t\}} S_{v_i^t t_j^t}, \tag{10}$$

where $j^*$ is the index of the candidate matching text $t_{j^*}^t$. Similarly, the candidate matching text $t_{j^*}^t$ is further calculated back to video set in a similar way, and the corresponding candidate matching video $v_{i^*}^t$ can be nominated according to $\mathbf{S}_{t_j^t}^{\mathcal{T}\to\mathcal{V}}$. As in Fig. 4, this dual mapping and matching operation between $\mathcal{V}^t$ and $\mathcal{T}^t$ determines the dual alignment consistency as:

$$\begin{cases} \left(v_{i^*}^t, t_{j^*}^t\right) \text{is an aligned pair}, & \text{if} \quad v_{i^*}^t = v_i^t, \\ \left(v_{i^*}^t, t_{j^*}^t\right) \text{is a misaligned pair}, & \text{if} \quad v_{i^*}^t \neq v_i^t. \end{cases} \tag{11}$$

The dual alignment consistency requires $v_{i^*}^t$ and $t_{j^*}^t$ to be the reciprocal nearest neighbor of each other, indicating a truly aligned (or positive) pair. To that end, we can obtain $n_p$ positive pairs in one batch from target dataset, denoted as $\left\{(\mathcal{V}^p, \mathcal{T}^p) = (v_i^p, t_i^p)_{i=1}^{n_p}\right\}$, where $(v_i^p, t_i^p) = \left(v_{i^*}^t, t_{j^*}^t\right)$.

To further boost the accuracy of aligned pairs, we introduce $T$ as the threshold to sort the similarities of all pairs in one batch with descending order and choose the $T$-th value, implying to select top $T$ similar pairs. The intuition is that a truly positive video-text pair should not only be the most similar to each other, but also have a relatively high similarity score compared with all the misaligned pairs. We also conduct some vanilla aligning mechanisms and ablations on threshold $T$, which are reported in Sec. 4 (Tab. 3 and Fig. 5).

With these self-discovered matching pairs, we can treat the pairwise misalignment issue as a fully supervised problem to benefit the model training. Similar to $\mathcal{L}_{\mathcal{S}}$ in source domain, the dual alignment consistency loss can be defined as:

$$\begin{aligned} \mathcal{L}_{\mathcal{P}} = -\frac{1}{2}\Big(log\frac{exp\left(s\left(v_i^p, t_i^p\right)/\tau\right)}{\sum_j^{n_p} exp\left(s\left(v_i^p, t_j^p\right)/\tau\right)} \\ + log\frac{exp\left(s\left(t_i^p, v_i^p\right)/\tau\right)}{\sum_j^{n_p} exp\left(s\left(t_i^p, v_j^p\right)/\tau\right)}\Big) \end{aligned} \tag{12}$$

During the training process, the positive pairs are increasing progressively and the noisy ones will potentially be aligned in the later stages (empirical results can be found in Fig. 5). To this end, more target samples are truly aligned as relevant video-text pairs, generating discriminative features in target domain.

### 3.4. Overall Training

In a nutshell, we minimize the sum of the above losses, including the semantic embedding loss $\mathcal{L}_{\mathcal{S}}$ in source domain, the domain adaptation loss $\mathcal{L}_{\mathcal{D}}$ for alleviating domain shift, and the dual alignment consistency loss $\mathcal{L}_{\mathcal{P}}$ for aligning positive pairs in target domain. Combining these loss terms together, the overall objective function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{S}} + \lambda_1 \mathcal{L}_{\mathcal{D}} + \lambda_2 \mathcal{L}_{\mathcal{P}}, \tag{13}$$

where $\lambda_1$ and $\lambda_2$ are hyper parameters for balancing these terms. The parameters of the whole network can readily be updated by the stochastic gradient descent (SGD) algorithm and the chain rules.

## 4. Experiments

### 4.1. Experimental Setting

**Datasets.** In this paper, we take advantage of existing datasets across three domains to explore the UDAVR task. To be specific, we construct a comprehensive evaluation benchmark which is the combination of three widely used datasets, i.e., MSR-VTT (Mt) [61], MSVD (Md) [23] and TGIF (Tf) [34]. An overview of three datasets is given in Tab. 1. The diversities of different datasets, e.g., lengths, numbers and video scenes, contribute to the domain shift in UDAVR task.

**Evaluation Metrics.** We adopt standard retrieval metrics (following [10, 25, 28]) to evaluate the performance of video-text retrieval. We measure rank-based performance by R@K (higher is better) and also report Median Rank, i.e., MR, (lower is better).

**Implementation Details.** For fair comparison, we utilize the same architecture of the video and text encoders as in GPO [5], which is the state-of-the-art baseline for video-text retrieval. Note that our method is orthogonal to the visual and textual encoder, allowing us to flexibly embrace state-of-the-art visual and textual encoders, of which the details are discussed in ablation studies. The length of shared embedding $M$ is set to 1024. Moreover, we adopt the Adam optimizer for all our experiments, set $\lambda_1$=$\lambda_2$ = 0.1. We set the mini-batch size to 64, and utilize a step-decayed learning rate with initialization value 0.0001.

Table 2. Effect of $\mathcal{L}_{\mathcal{P}}$ and $\mathcal{L}_{\mathcal{D}}$.

| Method | Tf→Mt | | | Tf→Md | | |
|---|---|---|---|---|---|---|
| | R1↑ | R10↑ | MR↓ | R1↑ | R10↑ | MR↓ |
| DADA(w/o $\mathcal{L}_{\mathcal{P}}$) | 2.98 | 15.93 | 97 | 10.26 | 39.18 | 18 |
| DADA(w/o $\mathcal{L}_{\mathcal{D}}$) | 4.12 | 20.29 | 75 | 11.47 | 41.90 | 17 |
| **DADA** | **5.30** | **24.54** | **50** | **14.34** | **48.77** | **11** |

Table 3. Comparison of different alignment mechanisms.

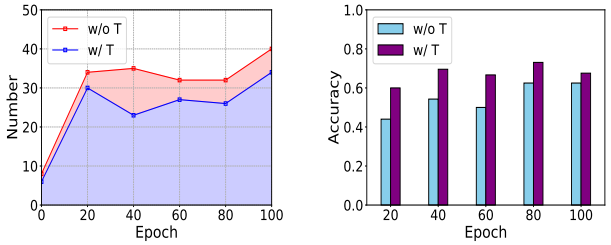
| Method | Tf→Mt | | | Tf→Md | | |
|---|---|---|---|---|---|---|
| | R1↑ | R10↑ | MR↓ | R1↑ | R10↑ | MR↓ |
| DADA(w/ text) | 3.52 | 17.29 | 111 | 13.16 | 45.67 | 13 |
| DADA(w/ video) | 3.26 | 15.61 | 146 | 13.04 | 46.31 | 13 |
| DADA(w/o T) | 4.84 | 22.47 | 67 | 13.53 | 44.77 | 15 |
| **DADA** | **5.30** | **24.54** | **50** | **14.34** | **48.77** | **11** |

## 4.2. Ablation Studies

**Effect of $\mathcal{L}_{\mathcal{P}}$ and $\mathcal{L}_{\mathcal{D}}$.** To evaluate the contribution of $\mathcal{L}_{\mathcal{P}}$ and $\mathcal{L}_{\mathcal{D}}$, we train the model by removing each component solely and present the results in Tab. 2. The results of DADA(w/o $\mathcal{L}_{\mathcal{D}}$) and DADA(w/o $\mathcal{L}_{\mathcal{P}}$) are inferior to the full DADA method, verifying the effectiveness of both components. Besides, DADA(w/o $\mathcal{L}_{\mathcal{P}}$) achieves worse performance than DADA(w/o $\mathcal{L}_{\mathcal{D}}$), further indicating that the dual alignment consistency is more important than simply alleviating the domain shift.

**Effect of Dual Alignment Consistency.** To explore the effectiveness of Dual Alignment Consistency (DAC), we comprehensively investigate several alignment mechanisms and show the results in Tab. 3. Specifically, DADA (w/ text) selects the unique text with the highest similarity for each target video. Similarly, DADA (w/ video) selects the unique video with the highest similarity for each target text. DADA (w/o $T$) removes the threshold $T$ of DAC. The results of DADA (w/ text) and DADA (w/ video) are worse than the full DADA method, demonstrating that the dual alignment mechanism is superior to aligning from only one modality stream. Meanwhile, the result of DADA (w/o $T$) is also inferior, which proves that the constraint on high similarities of truly aligned pairs is effective.

Furthermore, Fig. 5 shows the effect of threshold $T$ in dual alignment consistency within one batch during the training procedure. We can find that in Fig. 5(a), the number of dual aligned pairs (w/ $T$) is relatively smaller than that of w/o $T$. This is acceptable since the threshold $T$ constrains that aligned pairs should also have relatively high similarities compared to all pairs. As in Fig. 5(b), however, the accuracy of truly aligned pairs (w/ $T$) is evidently higher than that of w/o $T$, indicating that considering the threshold $T$ of DAC ensures the number of truly aligned pairs. As the training proceeds, the number of truly aligned pairs increases adaptively, which justifies the intuition that positive pairs are increasing and noisy ones are diminished.




(a) (b)

Figure 5. Analysis on threshold $T$ of dual alignment consistency. As training epochs proceeds, (a) numbers of dual aligned pairs and (b) accuracy of truly aligned pairs in one batch.

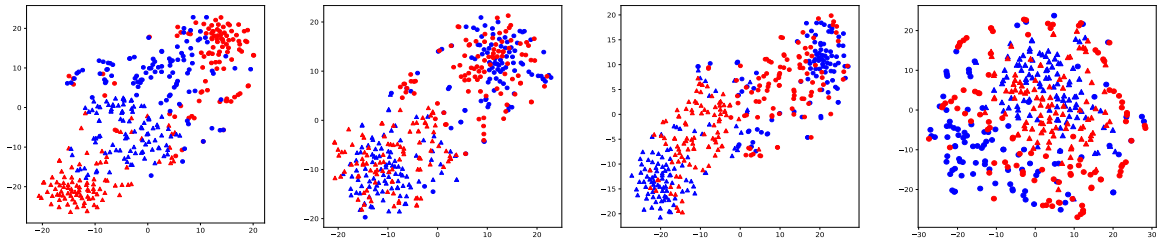
Table 4. Impact of different source datasets scales. Split-1/2/3 denotes 7,945/39,726/79,451 training data in source domain, respectively. Split-1 is adopted as the baseline.

| Splits | Tf→Mt | | | Tf→Md | | |
|---|---|---|---|---|---|---|
| | R1↑ | R10↑ | MR↓ | R1↑ | R10↑ | MR↓ |
| split-1 | 5.30 | 24.54 | 50 | 14.34 | 48.77 | 11 |
| split-2 | 7.82 | 31.38 | 33 | 19.10 | 54.02 | 8 |
| split-3 | 9.08 | 35.79 | 24 | 20.58 | 57.40 | 7 |

**Impact of source datasets scales.** To assess the impact of source datasets scales on the UDAVR task, we randomly split the TGIF dataset into three splits: Split-1/2/3 denotes 7,945/39,726/79,451 training data in source domain, respectively. The results are shown in Tab. 4. We observe that on one hand, increasing the number of training data in source dataset brings a trend of performance gain (split-1 to split-2). This is reasonable since large-scale source dataset usually provide more knowledge when transferring to the target domain. On the other hand, when the number goes to a relatively large value (split-2 to split-3), the performance gain is mostly marginal when transferring from TGIF to MSVD. We argue that the pairwise misalignment issue in UDAVR task can not be solved by simply increasing the number of source training data. Considering the data volume of MSR-VTT and MSVD, we adopt split-1 as the baseline for TGIF dataset.

**Generalization to different video-text retrieval methods.** As shown in Tab. 5, we implement several state-of-the-art video-text retrieval methods and the corresponding combinations with our DADA. Clearly, our method consistently improve the performances on target domain when combined with original baselines. Surprisingly, when combined with CLIP based methods, our DADA can still contribute to a remarkable performance gain. This verifies that the pairwise misalignment issue can't be diminished by merely adopting more powerful cross-modal retrieval methods, justifying the efficacy of our method.

**Sensitivity of Hyper-parameters.** We conduct experiments under the setting of Tf→Mt and Tf→Md, and present



(a) Source Only (b) MMD [40] (c) ACP [39] (d) Our method

Figure 6. The t-SNE visualizations of (a) Source Only, (b) MMD, (c) ACP and (d) Our method. Blue/red denotes source/target domain, while circles/triangles denote videos/texts. Our method progressively generates truly aligned video-text pairs in target domain, i.e., red circles and triangles are close together if they are semantically relevant (best viewed in color).

Table 5. Generalization to different video-text retrieval methods: (a) Single feature based methods (b) Multi-feature based methods and (c) CLIP based methods.

| | Backbone | Tf→Mt | | | Tf→Md | | |
|---|---|---|---|---|---|---|---|
| | | R1↑ | R10↑ | MR↓ | R1↑ | R10↑ | MR↓ |
| (a) | HGR [10] | 2.20 | 11.98 | 154 | 9.25 | 37.73 | 21 |
| | **HGR + DADA** | **3.82** | **18.75** | **96** | **10.97** | **39.16** | **20** |
| | GPO [5] | 2.69 | 13.63 | 144 | 9.39 | 37.77 | 20 |
| | **GPO + DADA** | **5.30** | **24.54** | **50** | **14.34** | **48.77** | **11** |
| (b) | CE [37] | 2.93 | 14.7 | 122 | 10.3 | 39.2 | 18 |
| | **CE + DADA** | **5.67** | **25.30** | **47** | **15.20** | **49.87** | **11** |
| | MMT [21] | 4.20 | 22.30 | 78 | 12.45 | 46.53 | 18 |
| | **MMT + DADA** | **6.23** | **27.31** | **38** | **17.30** | **50.34** | **10** |
| (c) | CLIP4CLIP [44] | 7.20 | 28.50 | 35 | 19.23 | 50.23 | 8 |
| | **CLIP4CLIP + DADA** | **8.62** | **33.12** | **28** | **21.30** | **58.90** | **6** |
| | CLIP2Video [19] | 7.80 | 31.50 | 31 | 20.23 | 51.20 | 8 |
| | **CLIP2Video + DADA** | **8.90** | **36.50** | **25** | **23.10** | **63.6** | **5** |

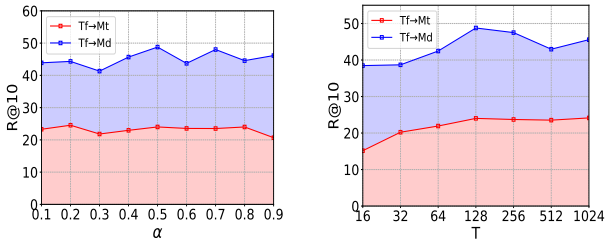



Figure 7. Sensitivity of hyper-parameters.

the sensitivity of hyper-parameters $\alpha$ and $T$ in Fig. 7. Within a wide range of $\alpha$ in [0.1, 0.9], the performance only varies in a small range, indicating the robustness to different choices of $\alpha$. Similarly, when progressively increasing threshold $T$ in DAC from 16 to 1,024, our method consistently performs well and achieves the best when $T = 128$. Thus, we set $\alpha$ to 0.7 and $T$ to 128 under all settings.

**Feature Visualisations.** We randomly choose 100 pairs in source and target domain respectively, and show the t-SNE [57] visualizations of Source Only, MMD, ACP and our method in Fig. 6. As can be seen, (a) Source Only shows that blue and red features are clearly separated, indicating the video and text domain shifts. (b) MMD and (c) ACP alleviate the domain shift, whereas inevitably mix up target videos and texts, ignoring whether they are a relevant pair or not. Obviously, our method not only diminishes the domain shift, i.e., blue and red features are mixed up, but also generates aligned pairs in target domain, i.e., red circles and triangles are close together if relevant.

### 4.3. Comparison with the State-of-the-arts

We compare our method with several state-of-the-art baselines across three categories, i.e., Source Only, DA methods and UDAVR methods. As a lower bound, we include the non-adapted Source Only results, which directly applies the model trained on the source domain to the target domain. We also implement five classification-based (i.e. typical) DA methods and modify them for the UDAVR task, including MMD [40], CORAL [53], DANN [22], IDM [14] and SCDA [33]. Moreover, we compare to three recently proposed UDAVR methods, i.e., MAN [17], CAPQ [9] and ACP [39]. For fairness, all methods adopt the same features and the backbone network as [5].

Tab. 6 shows that: (1) As the lower bound, Source Only achieves the worst performance, which identifies the existed domain shift problem. (2) Traditional DA methods in setting (a) are ineffective for the challenging UDAVR task, which can only slightly outperform the Source Only baseline. We owe this to that there exists no identical label set in UDAVR, which is the key difference of classification tasks. (3) Our method consistently outperforms other UDAVR methods in setting (b) on all adaptation directions across the three datasets, which demonstrates the effectiveness of dual alignment consistency. Compared with the SOTA method ACP, DADA achieves 20.18% and 18.61% relative improvements on R@1 under the setting of Tf→Mt and Tf→Md respectively.

Table 6. Comparison with different baselines. We denote Tf the TGIF, Mt the MSR-VTT and Md the MSVD dataset.

| | Method | Tf→Mt | | | Mt→Tf | | | Tf→Md | | | Md→Tf | | | Mt→Md | | | Md→Mt | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R1↑ | R10↑ | MR↓ | R1↑ | R10↑ | MR↓ | R1↑ | R10↑ | MR↓ | R1↑ | R10↑ | MR↓ | R1↑ | R10↑ | MR↓ | R1↑ | R10↑ | MR↓ |
| | Source Only | 2.69 | 13.63 | 144 | 6.30 | 25.43 | 60 | 9.39 | 37.77 | 20 | 3.80 | 16.99 | 102 | 15.02 | 46.96 | 12 | 2.50 | 13.27 | 136 |
| (a) | MMD [40] | 2.68 | 13.59 | 135 | 6.77 | 27.11 | 54 | 9.11 | 36.11 | 23 | 3.50 | 16.28 | 119 | 15.31 | 47.65 | 12 | 2.62 | 13.18 | 136 |
| (a) | CORAL [53] | 2.74 | 14.07 | 128 | 6.56 | 26.49 | 52 | 9.44 | 37.87 | 21 | 3.65 | 17.34 | 108 | 15.65 | 49.43 | 11 | 2.65 | 13.34 | 138 |
| (a) | DANN [22] | 2.76 | 13.94 | 127 | 6.86 | 27.17 | 48 | 9.27 | 38.00 | 20 | 3.74 | 16.72 | 103 | 15.67 | 48.67 | 11 | 2.62 | 13.17 | 134 |
| (a) | IDM [14] | 2.59 | 13.11 | 149 | 7.12 | 25.35 | 60 | 8.05 | 35.51 | 23 | 3.24 | 15.78 | 120 | 13.96 | 47.77 | 12 | 2.54 | 12.39 | 165 |
| (a) | SCDA [33] | 2.79 | 14.22 | 130 | 6.92 | 26.70 | 53 | 9.84 | 37.11 | 22 | 3.30 | 17.02 | 108 | 15.64 | 48.65 | 11 | 2.55 | 12.98 | 138 |
| (b) | MAN [17] | 2.53 | 12.98 | 144 | 6.42 | 25.96 | 63 | 8.84 | 37.06 | 21 | 3.06 | 16.31 | 119 | 15.05 | 48.51 | 11 | 2.40 | 12.00 | 174 |
| (b) | CAPQ [9] | 3.46 | 17.02 | 110 | 7.33 | 25.64 | 62 | 9.30 | 37.97 | 21 | 3.97 | 17.75 | 113 | 15.66 | 49.08 | 11 | 3.35 | 15.47 | 158 |
| (b) | ACP [39] | 4.41 | 21.72 | 64 | 7.83 | 26.72 | 50 | 12.09 | 41.38 | 18 | 5.12 | 21.46 | 82 | 17.87 | 54.34 | 8 | 5.90 | 25.68 | 54 |
| Ours | **DADA** | **5.30** | **24.54** | **50** | **8.21** | **28.97** | **45** | **14.34** | **48.77** | **11** | **6.03** | **22.52** | **78** | **18.97** | **57.93** | **7** | **6.40** | **27.61** | **42** |


Figure 8. Qualitative results of query texts and corresponding videos along with the changes in rank A←B←C, where A denotes the rank of DADA, B the ACP method and C the Source Only.

### 4.4. Qualitative Results

As in Fig. 8, given a query text, we present how the rank of the relevant video changes with different methods. Obviously, our method results in higher ranks of relevant videos compared with Source Only and ACP. Interestingly, our method performs worse given the query '*A fox is diving into the snow*', which might be owed to the confusion of the white fox and the background.

## 5. Conclusion

In this paper, we focus on the notoriously challenging task, i.e., UDA Video-text Retrieval (UDAVR), and develop the simple yet effective Dual Alignment Domain Adaptation (DADA) method. We introduce the cross-modal semantic embedding and domain adaptation to simultaneously generate discriminative source features and alleviate the video and text domain shifts. To tackle the *pairwise alignment* issue, we propose the Dual Alignment Consistency (DAC), which progressively generates truly aligned target pairs and ensures the discriminability of target features. Extensive experiments justify our superiority.

## 6. Acknowledgement

This work was supported by the National Key R&D Program of China under Grant 2022YFB3103500, the National Natural Science Foundation of China under Grants 62106258, 62006242 and 62202459, and the China Postdoctoral Science Foundation under Grant 2022M713348 and 2022TQ0363, and Young Elite Scientists Sponsorship Program by BAST (NO.BYESS2023304).

## References


[1] Nakul Agarwal, Yi-Ting Chen, Behzad Dariush, and Ming-Hsuan Yang. Unsupervised domain adaptation for spatio-temporal action localization. In *British Machine Vision Conference*, 2020. 2

[2] Samuel Albanie, Yang Liu, Arsha Nagrani, Antoine Miech, Ernesto Coto, Ivan Laptev, Rahul Sukthankar, Bernard Ghanem, Andrew Zisserman, Valentin Gabeur, Chen Sun, Karteek Alahari, Cordelia Schmid, Shizhe Chen, Yida Zhao, Qin Jin, Kaixu Cui, Hui Liu, Chen Wang, Yudong Jiang, and Xiaoshuai Hao. The end-of-end-to-end: A video understanding pentathlon challenge (2020). *arXiv preprint arXiv:2008.00744*, 2020. 2

[3] Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*. 2

[4] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5716–5725, 2018. 2, 4

[5] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798, 2021. 2, 3, 5, 7

[6] Min-Hung Chen, Zsolt Kira, Ghassan Alregib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6320–6329, 2019. 2

[7] Min-Hung Chen, Baopu Li, Yingze Bao, and Ghassan AlRegib. Action segmentation with mixed temporal domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 2

[8] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[9] Qingchao Chen, Yang Liu, and Samuel Albanie. Mind-the-gap! unsupervised domain adaptation for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1072–1080, 2021. 2, 7, 8

[10] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10635–10644, 2020. 2, 3, 5, 7

[11] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 521–530, 2017. 2, 4

[12] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, 2020. 2

[13] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 1

[14] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. Idm: An intermediate domain module for domain adaptive person re-id. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 4, 7, 8

[15] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 467–483, 2018. 1, 2

[16] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9346–9355, 2019. 3

[17] Jianfeng Dong, Zhongzi Long, Xiaofeng Mao, Changting Lin, Yuan He, and Shouling Ji. Multi-level alignment network for domain adaptive cross-modal retrieval. *Neurocomputing*, 440:207–219, 2021. 2, 7, 8

[18] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image CLIP. *arXiv preprint arXiv:2106.11097*, 2021. 2

[19] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image CLIP. *CoRR*, abs/2106.11097, 2021. 7

[20] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, 2020. 2

[21] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229, 2020. 7

[22] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning*, pages 1180–1189, 2015. 2, 7, 8

[23] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2712–2719, 2013. 5

[24] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In *European Conference on Computer Vision*, 2022. 2

[25] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Listen and look: Multi-modal aggregation and co-attention network for video-audio retrieval. In *International Conference on Multimedia and Expo*, pages 1–6, 2022. 5

[26] Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, and Weiping Wang. In *International Conference on Multimedia Retrieval*, pages 135–143, 2021. 2

[27] Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, Weiping Wang, and Dan Meng. What matters: Attentive and relational feature aggregation network for video-text retrieval. In *International Conference on Multimedia and Expo*, pages 1–6, 2021. 2

[28] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multimodal data augmentation. In *Winter Conference on Applications of Computer Vision Workshops*, pages 379–389, 2023. 5

[29] Xin Huang and Yuxin Peng. Deep cross-media knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8837–8846, 2018. 2

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, pages 4904–4916, 2021. 3

[31] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C. H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4943–4953, 2022. 2

[32] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C. H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4943–4953, 2022. 3

[33] Shuang Li, Mixue Xie, Fangrui Lv, Chi Harold Liu, Jian Liang, Chen Qin, and Wei Li. Semantic concentration for domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9102–9111, 2021. 1, 2, 7, 8

[34] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 5

[35] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 1

[36] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*, pages 2713–2724, 2020. 1

[37] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *British Machine Vision Conference*, 2019. 2, 7

[38] Yu Liu, Huai Chen, Lianghua Huang, Di Chen, Bin Wang, Pan Pan, and Lisheng Wang. Animating images to transfer CLIP for video-text retrieval. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1906–1911, 2022. 2

[39] Yang Liu, Qingchao Chen, and Samuel Albanie. Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14954–14964, 2021. 2, 7, 8

[40] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning*, 2015. 1, 2, 7, 8

[41] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the International Conference on Machine Learning*, 2017. 1, 2

[42] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, and Shilei Wen. Attention clusters: Purely attention based local feature integration for video classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1

[43] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of CLIP for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2

[44] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, pages 293–304, 2022. 7

[45] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval. In *ACM International Conference on Multimedia*, pages 638–647, 2022. 2

[46] Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. Query and keyframe representations for ad-hoc video search. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*, pages 407–411, 2017. 1

[47] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 2

[48] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roychowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 2

[49] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 119–129, 2020. 2

[50] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11815–11822, 2020. 1

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 3

[52] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9787–9795, 2021. 1

[53] Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, volume 9915, pages 443–450, 2016. 1, 7, 8

[54] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 1

[55] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1456–1465, 2019. 1

[56] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1, 2

[57] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008. 7

[58] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016. 2

[59] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2VLAD: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[60] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9092–9101, 2021. 2

[61] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016. 5

[62] Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, and Kai Lei. Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia*, 21(4):1047–1061, 2018. 2, 4

[63] Jin Yuan, Zhengjun Zha, Yantao Zheng, Meng Wang, Xiangdong Zhou, and Tatseng Chua. Utilizing related samples to enhance interactive concept-based video search. *IEEE Transactions on Multimedia*, pages 1343–1355, 2011. 1

[64] Wentian Zhao, Xinxiao Wu, and Jiebo Luo. Cross-domain image captioning via cross-modal retrieval and model adaptation. *IEEE Transactions on Image Processing*, 30:1180–1192, 2020. 2, 4