Unsupervised Domain Adaptative Temporal Sentence Localization with Mutual Information Maximization

Daizong Liu[1*], Xiang Fang[2*], Xiaoye Qu[3], Jianfeng Dong[4], He Yan[5], Yang Yang[6], Pan Zhou[3†], Yu Cheng[7]

[1]Wangxuan Institute of Computer Technology, Peking University
[2]Nanyang Technological University
[3]Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science of Technology
[4]College of Computer Science and Technology, Zhejiang Gongshang University
[5]Protagolabs Inc.
[6]Meta Platforms Inc.
[7]Department of Computer Science and Engineering, The Chinese University of Hong Kong
dzliu@hust.edu.cn, xfang9508@gmail.com, xiaoye@hust.edu.cn, dongjf24@gmail.com, he.yan@protagolabs.com, yang.angela06@gmail.com, panzhou@hust.edu.cn, chengyu@cse.cuhk.edu.hk

Abstract

Temporal sentence localization (TSL) aims to localize a target segment in a video according to a given sentence query. Though respectable works have made decent achievements in this task, they severely rely on abundant yet expensive manual annotations for training. Moreover, these trained data-dependent models usually can not generalize well to unseen scenarios because of the inherent domain shift. To facilitate this issue, in this paper, we target a practical but challenging setting: unsupervised domain adaptative temporal sentence localization (UDA-TSL), which explores whether the localization knowledge can be transferred from a fully-annotated data domain (source domain) to a new unannotated data domain (target domain). Particularly, we propose an effective and novel baseline for UDA-TSL to bridge the multi-modal gap across different domains and learn the potential correspondence between the video-query pairs in target domain. We first develop separate modality-specific domain adaptation modules to smoothly balance the minimization of the domain shifts in cross-dataset video and query domains. Then, to fully exploit the semantic correspondence of both modalities in target domain for unsupervised localization, we devise a mutual information learning module to adaptively align the video-query pairs which are more likely to be relevant in target domain, leading to more truly aligned target pairs and ensuring the discriminability of target features. In this way, our model can learn domain-invariant and semantic-aligned cross-modal representations. Three sets of migration experiments show that our model achieves competitive performance compared to existing methods.

Introduction

Temporal sentence localization (TSL) (Anne Hendricks et al. 2017; Gao et al. 2017) is an important yet challenging task in video understanding, which has drawn increasing attention due to its vast potential applications, such as activity detection (Dong et al. 2022a) and human-computer interaction (Dong et al. 2022b). Given an untrimmed video, it aims to retrieve a temporal video segment that semantically corresponds to a given sentence query.

Most previous TSL works (Yuan et al. 2019; Zhang et al.

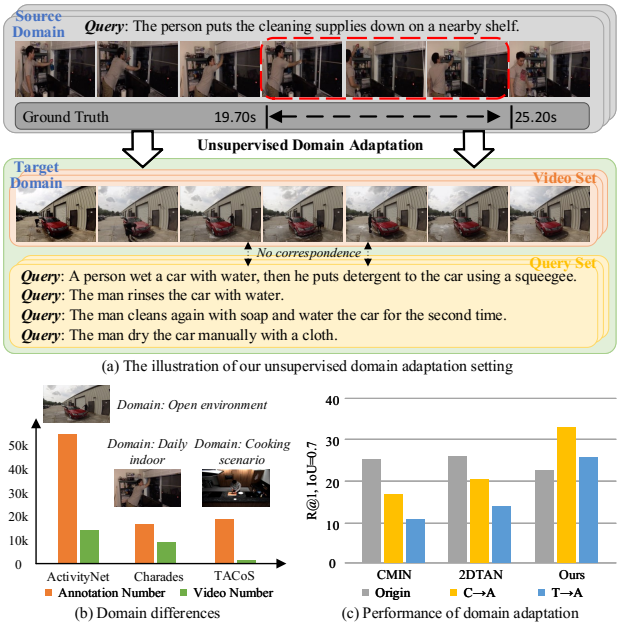

Figure 1: (a) We investigate UDA-TSL, which transfers the localization knowledge from a fully-annotated source domain to a new unannotated target domain. (b) The inherent domain differences (including number of annotated moments, number of videos, and video content types) among three typical datasets for TSL task. (c) The domain adaptation performances of existing methods are limited.

*The two authors contributed equally.
†Corresponding Author.

2020a; Chen et al. 2020a; Liu et al. 2020, 2021b, 2022a, 2021a, 2022c, 2023a; Liu, Qu, and Zhou 2021; Liu, Qu, and Hu 2022) are under fully-supervised setting, where each frame is manually labeled as the query-relevant or query-irrelevant frame. Instead of using such dense frame annotations, some recent works try to explore a weakly-supervised setting (Lin et al. 2020; Zhang et al. 2020c; Liu et al. 2022b; Chen et al. 2020c; Song et al. 2020) with only the video-query correspondence to alleviate the reliance to a certain extent. Despite their great advances, the above two types of supervised methods severely rely on abundant video-query annotations, which is both time-consuming and labor-intensive to collect in real-world scenarios. Although few works (Liu et al. 2022b; Fang et al. 2023a) try to design TSL model without any supervision, their performances are validated to be less satisfied.

In this paper, considering that we always can collect a few fully-annotated datasets and a massive number of unannotated datasets in practice, we make the attempt to explore whether a TSL model can learn more generalizable localization knowledge from a known dataset and perform well on an unseen dataset. In this manner, we can alleviate the reliance of the video-query correspondence of the new dataset, and directly utilize the knowledge of previously annotated data to fit it. To this end, we propose a more practical but challenging scenario for TSL task, *i.e.*, unsupervised domain adaptative temporal sentence localization (UDA-TSL), which has the fully annotation in source domain and no annotation in target domain as shown in Figure 1 (a). Different from previous TSL settings, this new UDA-TSL suffers from two major issues: (1) *How to transfer the knowledge from the source domain into the target domain?* As shown in Figure 1 (b), the widely used TSL datasets (ActivityNet, Charades, TACoS) contain different numbers of annotation video-query pairs and training samples. Besides, the domains of their video content focus on different real-world scenarios with different objects and backgrounds. Therefore, previous TSL models trained on label-rich source domain usually can not generalize well to another label-scare target domain due to these domain shifts among data distribution as shown in Figure 1 (c), severely limiting their practical applications. (2) *How to deal with the unknown video-query correspondence in the unsupervised target domain?* Since the target domain only contains the unlabeled video and query sets, it is also important to mine their potential correspondence for correlating the matched video-query pair. Once we obtain the positive matched multi-modal pair, we can transfer the knowledge of the labeled source domain to train the discriminative multi-modal representations in the target domain for learning the possible segment localization.

To tackle these issues, we propose a novel baseline model for UDA-TSL to bridge the multi-modal gap across different domains and learn the potential correspondence between the video-query pairs in target domain. Specifically, given the video and query sets from both source and target domains, we first utilize the same video and query encoders to generate corresponding feature representations. Since different domains usually have inconsistent data distributions leading to the domain shift, we then develop a modality-specific domain adaptation module to relieve such divergence by optimizing an appropriate intermediate domain to best bridge the source and target domains. To further mine the possible video-query correspondence in the unsupervised target domain for latter localization, we devise an effective Mutual Information Maximization (MIM) module to capture the cross-modal semantical relevance in bi-directional video-to-query and query-to-video ways. During the learning, MIM adaptively aligns the video-query pairs which are more likely to be relevant in target domain, enabling that positive pairs are increasing progressively and the discriminability of target features is generalized like the source statistics. During the inference, we directly utilize the learned query-related frame-wise scores for localizing the interested segment. To sum up, our main contributions are:

- In this paper, we tackle a more practical but challenging TSL setting, called UDA-TSL, which transfers the localization knowledge from a fully-annotated source domain to a new unannotated target domain. Different from previous TSL works, our new setting can be applied to real-world scenarios for addressing the online learning problem and unseen localization cases.
- We propose an effective modality-specific domain adaptation module to reduce the domain gaps between the video/query sets in different domains by optimizing an appropriate intermediate domain. To learn the possible video-query correspondence in the unsupervised target domain, we also develop a MIM module to mine the truly positive multimodal pairs for self-supervising the discriminative representation learning.
- We conduct the UDA experiments on three widely used TSL datasets (ActivityNet Captions, Charades-STA, and TACoS). Extensive results show that our proposed model performs much better than existing approaches.

Related Work

Temporal sentence localization. Most of the existing TSL methods refer to fully-supervised setting where all video-query pairs are annotated in details, including corresponding segment boundaries. Therefore, the main challenge in such setting is how to align multi-modal features well to predict precise boundary. Some works (Gao et al. 2017; Zhang et al. 2019; Yuan et al. 2019; Zhang et al. 2020b; Chen et al. 2018; Qu et al. 2020; Liu and Hu 2022; Fang et al. 2022, 2023c, 2020; Fang and Hu 2020; Fang et al. 2021a,b; Liu et al. 2022d, 2023b; Fang et al. 2023b; Zheng et al. 2023; Zhu et al. 2023; Liu et al. 2023c,d) integrate sentence information with each fine-grained video clip unit, and predict the scores of candidate segments by gradually merging the fusion feature sequence over time. Without using proposals, some latest methods (Nan et al. 2021; Zhang et al. 2020a; Chen et al. 2020a) are proposed to leverage the interaction between video and sentence to directly predict the starting and ending frames. However, the above methods heavily rely on the datasets that require numerous manually labelled annotations for training. To ease the human labelling efforts, several recent works (Chen et al. 2020c; Song et al. 2020;

Lin et al. 2020; Zhang et al. 2020c) consider a weakly-supervised setting which only accesses the information of matched video-query pairs without accurate segment boundaries. However, their performance is less satisfied.

Unsupervised domain adaptation. Unsupervised domain adaptation (UDA) aims to transfer predictive models trained on fully-labeled data from a source domain to an unlabeled target domain. The primary objective of existing UDA methods, which are predominantly classification-based, is to mitigate the domain shift that occurs between the source and target domains (Qu et al. 2019; Damodaran et al. 2018; Ganin and Lempitsky 2015; Long et al. 2015, 2017; Tzeng et al. 2014). Furthermore, the realm of UDA has witnessed significant advancements in the domain of video-based tasks, such as video action recognition (Chen et al. 2019; Choi et al. 2020; Munro and Damen 2020) and video segmentation (Chen et al. 2020b,b). UDA techniques have been successfully extended to these video-related tasks, enabling knowledge transfer from the labeled source domain to the unlabeled target domain. Additionally, recent research efforts have explored the application of UDA in cross-modal tasks (*e.g.*, image captioning (Chen et al. 2017; Yang et al. 2018; Zhao, Wu, and Luo 2020), visual question answering (Chao, Hu, and Sha 2018), and image-text retrieval (Huang and Peng 2018)), where different modalities such as images and texts are involved. The most similar work to our work is (Hao et al. 2023), different from it, we make novel UDA designs in the specific TSL task.

The Proposed Method

Overview

Problem definition. Given an untrimmed video $V = \{v_i\}_{i=1}^{N_v}$ and corresponding language query $Q = \{q_j\}_{j=1}^{N_q}$, where N_v and N_q are the number of frames and words, traditional temporal sentence localization (TSL) task aims to localize the query-described activity segment from the video. In this paper, we investigate a more practical but challenging setting, called unsupervised domain adaptative TSL (UDA-TSL), which transfers the localization knowledge from the previous annotated source domain to a new unannotated target domain. In particular, we denote the source and target datasets as $\{\{V_I^s\}_{I=1}^{N_V^s}, \{Q_J^s\}_{J=1}^{N_Q^s}\}, \{\{V_l^t\}_{I=1}^{N_V^t}, \{Q_J^t\}_{J=1}^{N_Q^t}\}$, where s, t imply source and target domains, and N_V, N_Q denotes the numbers of videos and queries.

Pipeline. To tackle the UDA-TSL, we propose a novel framework as shown in Figure 2. After encoding the multi-modal features in different domains, to reduce the distribution shifts of the same modality across differnt domains, we propose to optimize an appropriate intermediate domain to better help the gradual adaptation between two domains. Compared to previous simple source-to-target domain closing paradigm, our strategy does not suffer from the huge domain shift and is able to smoothly transfer the knowledge across domains in a closest path. As for the discriminative representation learning, we train the source data with the collected pairwise video-query annotations. As for the unannotated target data, we devise a mutual information maximization module to align the representations of truely positive target video-text pairs which are more likely to be semantically relevant, and to avoid including the noisy ones which tend to be irrelevant. During the inference, we directly take the interacted query-related video representations for frame-wise scoring and segment determining.

Preparation

Video encoding. Following previous works (Zhang et al. 2019, 2020a,b), given the video $V \in \{V^s, V^t\}$ from any domain, we first extract its frame-wise features by a pre-trained C3D network (Tran et al. 2015), and then employ a multi-head self-attention (Vaswani et al. 2017) module to capture the long-range dependencies among video frames. We denote the extracted video features as $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^{N_v} \in \mathbb{R}^{N_v \times d}$, where d is feature dimension.

Query encoding. Similarly, given the query $Q \in \{Q^s, Q^t\}$ from any domain, we also follow previous works (Zhang et al. 2019, 2020a,b) to utilize the Glove (Pennington, Socher, and Manning 2014) embedding to encode each word into dense vector. We further employ the Bi-GRU (Chung et al. 2014) layers to encode the word-level sequential information in the whole sentence. The final word-level feature can be denoted as $\boldsymbol{Q} = \{\boldsymbol{q}_j\}_{j=1}^{N_q} \in \mathbb{R}^{N_q \times d}$.

Modality-Specific Domain Adaptation

Generally, different datasets usually have inconsistent data distributions and representations, thus leading to the domain shift problem (Na et al. 2021). To alleviate this issue, we propose to relieve the divergence of different statistics between source and target domains. Following previous UDA methods (Ganin et al. 2016; Hosseini-Asl et al. 2018; Huang, Peng, and Yuan 2018), a general way of reducing the domain discrepancy is to directly shift the source statistics close to the target one and vice versa. However, in practice, this strategy does not always work since there can be huge shift between the two extreme domains' distributions, as shown in Figure 2. Instead, there always exists an appropriate intermediate domain that is located along with the shift path to help the gradual adaptation between the two extreme domains. Once this intermediate domain is well determined, it is able to bridge the two extreme domains along which the source domain's knowledge can be smoothly transferred to guide the learning of the target domain. For example, if an intermediate domain is closer to the source domain, the source reliable labels can be more leveraged. On the contrary, the target domain's intrinsic distribution can be more exploited.

To achieve this, we first generate the initial yet coarse intermediate domain representations $\boldsymbol{V}^m$ following (Gopalan, Li, and Chellappa 2013), and define a function $P(\cdot)$ to represent the distribution of each modality data in each domain. Specifically, as for the video set $\{\boldsymbol{V}_I^s\}_{I=1}^{N_V^s}$ of source domain, we calculate its distribution as:

$$P(\boldsymbol{V}^s) = \Big(\frac{1}{N_V^s}\sum_{I=1}^{N_V^s}\big(\mu(\boldsymbol{V}_I^s) - \frac{\sum_{I=1}^{N_V^s}\mu(\boldsymbol{V}_I^s)}{N_V^s}\big)\Big)^{1/2}, \quad \text{where} \quad \mu(\boldsymbol{V}_I^s) = \frac{1}{N_v^s}\sum_{i=1}^{N_v^s}\boldsymbol{v}_i^s. \tag{1}$$

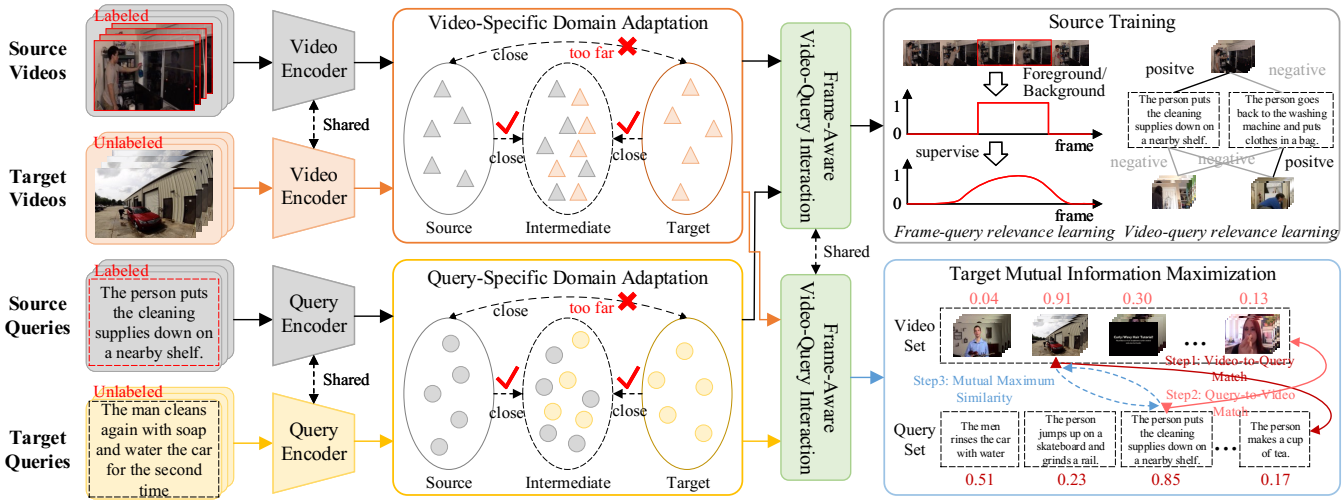

Figure 2: Overview of the proposed architecture for UDA-TSL task.

In this way, the distribution of target and intermediate domains can be represented as $P(\boldsymbol{V}^t), P(\boldsymbol{V}^m)$. To further tune the intermediate domain be located along the shortest geodesic path between the source and target domains, we introduce the domain factor α_1 for the source and target domains, which 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 two domains can be formulated as:

$$\frac{||P(\boldsymbol{V}^s) - P(\boldsymbol{V}^m)||_2^2}{||P(\boldsymbol{V}^t) - P(\boldsymbol{V}^m)||_2^2} = \frac{\alpha_1}{1-\alpha_1}. \tag{2}$$

Therefore, the video-specific domain shift problem can be converted into finding the appropriate intermediate domain by minimizing the intermediate domain loss as:

$$\mathcal{L}_V = \alpha_1||P(\boldsymbol{V}^t)-P(\boldsymbol{V}^m)||_2^2+(1-\alpha_1)||P(\boldsymbol{V}^s)-P(\boldsymbol{V}^m)||_2^2, \tag{3}$$

which guide the distribution of appropriate intermediate domain to keep the right distance to the source and target domains. Similarly, in the text stream, the query-specific domain adaptation loss can be computed as:

$$\mathcal{L}_Q = \alpha_2||P(\boldsymbol{Q}^t)-P(\boldsymbol{Q}^m)||_2^2+(1-\alpha_2)||P(\boldsymbol{Q}^s)-P(\boldsymbol{Q}^m)||_2^2. \tag{4}$$

The whole domain adaptation loss can be defined as:

$$\mathcal{L}_{domain} = \mathcal{L}_V + \mathcal{L}_Q. \tag{5}$$

Unsupervised Mutual Information Maximization

Note that in the UDA-TSL setting, there exists no identical label set for source and target domains, and the only supervision available is the semantic relationship in the source dataset. Although the modality-specific domain adaptation can alleviate the domain distribution shift, it is still not enough and hard for the model to learn the unannotated relationships between the video-query pair in the target domain. To this end, we propose an effective Mutual Information Maximization (MIM) module to explore the truly aligned target video-query pairs which are more likely to be semantically relevant, and to avoid including the noisy ones which tend to be irrelevant. By capturing the potential video-query correspondence via frame-wise score learning in the target domain, we can determine the query-related frames among the video for localization.

Specifically, given the target set $\{\{V_I^t\}_{I=1}^{N_V^t}, \{Q_J^t\}_{J=1}^{N_Q^t}\}$, we try to find if there exist truly positive video-text pairs (V_I^t, Q_J^t) can be considered as a truly positive pair if and only if V_I^t and Q_J^t are mutually the most semantically similar to each other. For a target video V_I^t, to calculate its relevance to a random query Q_J^t, we first interact their features as:

$$\begin{aligned} C_{I,J,i,j}^t &= \mathbf{w}^\top \tanh(\boldsymbol{W}_1 \boldsymbol{v}_{I,i}^t + \boldsymbol{W}_2 \boldsymbol{q}_{J,j}^t + \boldsymbol{b}_1), \\ \boldsymbol{s}_{I,J,i} &= \sum_{j=1}^{N_q} \text{softmax}(C_{I,J,i,j}^t) \cdot \boldsymbol{q}_{J,j}^t, \end{aligned} \tag{6}$$

where $\mathbf{W}_1^\alpha, \mathbf{W}_2^\alpha$ are projection matrices, $\mathbf{b}^\alpha$ is the bias and $\mathbf{w}^\top$ is the row vector (Zhang et al. 2019). $\boldsymbol{s}_{I,J,i}$ is the frame-aware query semantics. Then, we generate the query-relevant i-th frame features as:

$$\widetilde{\boldsymbol{v}}_{I,J,i}^t = \sigma(\boldsymbol{W}_3 \boldsymbol{s}_{I,J,i} + \boldsymbol{b}_2) \odot \boldsymbol{v}_{I,i}^t, \tag{7}$$

where σ is the sigmoid function, $\odot$ is the element-wise multiplication. The total query-to-video relevance can be formulated by:

$$r_{I,J} = \sum_{i=1}^{N_v} \text{softmax}(\text{MLP}_1(\widetilde{\boldsymbol{v}}_{I,J,i}^t)) \cdot \sigma(\text{MLP}_2(\widetilde{\boldsymbol{v}}_{I,J,i}^t)), \tag{8}$$

where the second item is the probability representing whether the i-th frame is relevant to the query or not, the first item is the normalized weight for aggregating all video frames. Therefore, we can obtain the similarities of V_I^t and all the target queries:

$$\boldsymbol{R}_{V_I^t} = [r_{I,1}, r_{I,2}, ..., r_{I,J}, ..., r_{I,N_Q^t}]. \tag{9}$$

Method	Charades→ActivityNet				ActivityNet→TACoS				TACoS→Charades			
	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
CBP	27.46	15.37	52.68	36.83	25.33	21.79	40.02	34.65	22.38	11.95	50.08	31.35
SCDM	28.02	15.84	52.51	34.16	22.68	17.45	38.56	31.90	35.95	25.18	52.69	39.83
CMIN	34.25	18.63	58.79	41.98	20.51	15.04	35.29	26.21	28.06	18.22	55.35	36.71
CSMGAN	36.92	20.04	62.46	49.57	29.63	18.07	49.32	40.15	36.45	22.86	61.73	40.50
2DTAN	39.17	21.76	69.33	57.50	33.72	21.16	54.84	42.39	25.81	17.37	57.29	33.48
DRN	41.39	24.27	71.42	43.87	32.07	19.96	47.54	31.28	36.16	24.52	66.90	42.64
MMN	44.06	24.98	72.25	58.19	36.94	22.08	55.73	46.81	33.73	20.04	62.42	44.29
NoUDA (Ours)	35.83	19.52	61.37	47.61	31.26	20.30	52.58	40.95	27.25	16.42	53.81	37.92
UDA (Ours)	**49.48**	**32.15**	**77.74**	**65.39**	**42.40**	**29.83**	**58.99**	**51.04**	**41.39**	**28.63**	**70.16**	**49.85**
Method	TACoS→ActivityNet				Charades→TACoS				ActivityNet→Charades			
	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
CBP	18.94	11.93	29.79	18.62	22.88	19.26	41.56	33.20	32.82	14.39	65.43	48.97
SCDM	19.65	11.80	30.51	17.27	18.97	16.82	35.05	28.41	52.56	34.82	72.75	56.41
CMIN	22.17	13.72	32.48	21.34	19.38	15.34	36.06	24.99	45.03	31.74	76.89	60.62
CSMGAN	23.88	14.67	39.07	27.30	25.43	16.12	44.96	36.72	45.60	32.28	77.53	54.27
2DTAN	24.90	16.38	45.04	30.15	30.12	19.81	51.29	39.87	36.34	22.61	74.14	48.56
DRN	24.93	18.52	41.74	27.44	28.60	16.73	43.00	28.15	50.47	29.02	86.25	57.19
MMN	28.29	20.86	47.27	32.63	34.09	19.17	52.08	43.24	50.78	23.17	79.53	56.80
NoUDA (Ours)	24.26	15.09	42.33	28.98	23.76	16.09	43.37	34.58	44.29	28.46	75.99	52.74
UDA (Ours)	**33.54**	**26.16**	**54.18**	**41.02**	**36.42**	**25.48**	**57.69**	**48.74**	**60.26**	**41.03**	**89.62**	**63.85**

Table 1: Performance comparison of existing SOTA methods on three widely used TSL datasets in the UDA setting.

After that, we choose the candidate matching query with maximum similarity via $argmax_{J^* \in \{1,2,\ldots,N_Q^t\}} r_{I,J^*}$, where J^* is the index of the candidate matching query. In turn, we utilize the candidate matching query $Q_{J^*}^t$ to further calculate back to the video set in a similar way, and obtain the corresponding candidate matching video $V_{I^*}^t$. We determine whether the unannotated video-query pair $(V_I^t, Q_{J^*}^t)$ is matched by:

$$\begin{cases} (V_{I^*}^t, Q_{J^*}^t) \text{ is matched}, & \text{if } V_{I^*}^t = V_I^t, \\ (V_{I^*}^t, Q_{J^*}^t) \text{ is not matched}, & \text{if } V_{I^*}^t \neq V_I^t. \end{cases} \tag{10}$$

This requires $V_I^t, Q_{J^*}^t$ to be the reciprocal nearest neighbor of each other, indicating a truly matched (or positive) pair.

With these self-discovered matching pairs, we take them as the positive samples and further construct other mismatched pairs as negative samples to train the localization model via contrastive learning as:

$$\begin{aligned} \mathcal{L}_C^t = &- \frac{1}{N_V^t} \sum_{I=1}^{N_V^t} \log \frac{\exp(r_{I,J^*}/\tau)}{\sum_{J=1, J\neq J^*}^{N_B} \exp(r_{I,J}/\tau)} \\ &- \frac{1}{N_Q^t} \sum_{J=1}^{N_Q^t} \log \frac{\exp(r_{I*,J}/\tau)}{\sum_{I=1, I\neq I^*}^{N_B} \exp(r_{I,J}/\tau)}, \end{aligned} \tag{11}$$

where τ is temperature parameter, N_B is batch size. In this manner, during the training process, the positive pairs are increasing progressively and the noisy ones will be eliminated, generating more discriminative features in target domain.

Training and Inference

Training. As for the source domain, since we have adequate annotations, we directly train the model with another contrastive loss $\mathcal{L}_C^s$ similar to Eq.(11) without MIM. To train corresponding localization results, we follow previous work (Zhang et al. 2020a) to generate foreground-background query-related frame-wise annotation $Y = \{y_i\}_{i=1}^{N_v}, y_i \in [0, 1]$, which supervises the frame-wise query-to-video relevance learning of Eq.(8) by cross-entropy loss as:

$$\mathcal{L}_{CE} = \text{CE}(y_i, \text{softmax}(\text{MLP}_1(\widetilde{\boldsymbol{v}}_{I,J,i}^t)) \cdot \sigma(\text{MLP}_2(\widetilde{\boldsymbol{v}}_{I,J,i}^t))). \tag{12}$$

By implementing the multi-modal domain adaptation losses and the target domain contrastive loss, the overall loss is:

$$\mathcal{L} = \mathcal{L}_{domain} + \mathcal{L}_C^t + \mathcal{L}_C^s + \mathcal{L}_{CE}. \tag{13}$$

Inference. During the inference, we feed each video-query pair of the target domain into encoders and then interact them to generate the frame-wise matching scores like Eq.(8). Following previous unsupervised work (Liu et al. 2022b), to predict the final localization result, we first locate the frame with the highest similarity score as the basic predicted segment, and add the left/right frames into the moment if the ratio of their scores to the frame score of the closest segment boundary is less than a certain threshold. In our all experiments, this threshold is set to 0.8 in ActivityNet Captions, TACoS and 0.9 in Charades-STA. We repeat this step to construct the moment until no frame can be added.

Experiments

Dataset

For fair comparison with existing TSL works, we utilize the same ActivityNet Caption (Caba Heilbron et al. 2015), TACoS (Regneri et al. 2013), and Charades-STA (Sigurdsson et al. 2016) datasets for evaluation. Specifically, ActivityNet Caption contains 20000 untrimmed videos with 100000 descriptions from YouTube. Following public split,

Method	Charades→ActivityNet				ActivityNet→TACoS				TACoS→Charades			
	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
w/o $\mathcal{L}_{domain}, \mathcal{L}_C^t$	35.83	19.52	61.37	47.61	31.26	20.30	52.58	40.95	27.25	16.42	53.81	37.92
w/o $\mathcal{L}_C^t$	40.65	24.22	68.01	55.36	35.87	24.36	55.58	44.26	33.50	21.26	59.82	43.12
w/o $\mathcal{L}_{domain}$	43.34	26.46	71.43	56.98	38.51	26.07	55.64	46.68	35.35	23.14	62.39	44.80
Full model	**49.48**	**32.15**	**77.74**	**65.39**	**42.40**	**29.83**	**58.99**	**51.04**	**41.39**	**28.63**	**70.16**	**49.85**
Method	TACoS→ActivityNet				Charades→TACoS				ActivityNet→Charades			
	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	R@1, IoU=0.3	R@1, IoU=0.5	R@5, IoU=0.3	R@5, IoU=0.5	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
w/o $\mathcal{L}_{domain}, \mathcal{L}_C^t$	24.26	15.09	42.33	28.98	23.76	16.09	43.37	34.58	44.29	28.46	75.99	52.74
w/o $\mathcal{L}_C^t$	28.29	18.88	46.25	32.43	28.12	20.44	49.32	41.48	51.37	33.93	81.91	56.32
w/o $\mathcal{L}_{domain}$	29.80	21.46	49.28	36.79	31.40	21.75	51.45	41.04	53.45	36.31	82.61	59.00
Full model	**33.54**	**26.16**	**54.18**	**41.02**	**36.42**	**25.48**	**57.69**	**48.74**	**60.26**	**41.03**	**89.62**	**63.85**

Table 2: Main ablation. We investigate the contribution of the domain adaptation module $\mathcal{L}_{domain}$ and the MIM module $\mathcal{L}_C^t$.

we use 37417, 17505, and 17031 sentence-video pairs for training, validation, and testing. TACoS contains 127 videos collected from cooking scenarios. We also follow the public split, which includes 10146, 4589, 4083 query-segment pairs for training, validation and testing. As for Charades-STA, there are 12408 and 3720 moment-query pairs in the training and testing sets, respectively.

Implementation Details

As for video encoding, following previous works (Zhang et al. 2020b; Wang et al. 2022), we apply the pre-trained C3D (Tran et al. 2015) model to encode the videos on ActivityNet Caption, TACoS, and VGG (Simonyan and Zisserman 2014) model on Charades-STA. Since some videos are overlong, we uniformly downsample the length of video feature sequences to $N_v = 200$ for ActivityNet Caption and TACoS datasets, $N_v = 64$ for Charades-STA dataset. As for sentence encoding, we set the length of word feature sequences to $N_q = 20$, and utilize Glove embedding (Pennington, Socher, and Manning 2014) to embed each word to 300 dimension features. The dimension d is set to 512. We train our model for 100 epochs with an early stopping strategy. Parameter optimization is performed by Adam optimizer with learning rate of 0.0005, linear decay rate of 1.0.

Comparison with State-of-the-Art

To evaluate our performance, we re-implement several state-of-the-art TSL methods for comparison: CBP (Wang, Ma, and Jiang 2020), SCDM (Yuan et al. 2019), CMIN (Zhang et al. 2019), CSMGAN (Liu et al. 2020), 2DTAN (Zhang et al. 2020b), DRN (Zeng et al. 2020), MMN (Wang et al. 2022). Note that, we implement all the methods on three datasets in the same UDA setting, *i.e.*, training on the source dataset and testing on the target dataset (source → target). As shown in Table 1, we find that all previous models achieve worse domain adaptation performance compared to their original results reported in their papers. We also test our baseline model NoUDA which does not use any domain adaptation strategy. It even performs much worse than previous works. This demonstrates that TSL models without UDA design will suffer from the domain gap and is quite limited in the UDA setting. Instead, our UDA method can achieve much better performance due to our new design for reducing the domain gap, demonstrating its effectiveness.

Component	Variant	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
MSDA	w/o DA	43.34	26.46	71.43	56.98
	w/ normal DA	46.99	30.39	75.65	62.30
	w/ our DA	**49.48**	**32.15**	**77.74**	**65.39**
MIM	w/o MIM	40.65	24.22	68.01	55.36
	w/ text	46.89	29.42	74.33	62.40
	w/ video	46.57	29.20	73.85	61.95
	w/ mutual	**49.48**	**32.15**	**77.74**	**65.39**

Table 3: Effect of the modality-specific domain adaptation module (MSDA) and the mutual information maximization module (MIM) on the Charades→ActivityNet task.

Ablation Study

Main ablation study. To demonstrate the effectiveness of each component in our model, we conduct ablation studies regarding the components (*i.e.*, modality-specific domain adaptation module $\mathcal{L}_{domain}$ and mutual information maximization module $\mathcal{L}_C^t$), and show the corresponding experimental results in Table 2. From this table, we can find that both modules contribute a lot to the final performances, demonstrating that they are able to reduce the domain gaps between the two datasets. Moreover, the MIM module $\mathcal{L}_C^t$ brings the largest improvement, demonstrating that it provides positive pseudo video-query paris for learning the discriminative frame-wise representations for accurate localization in the unsupervised target domain.

Effect of the modality-specific domain adaptation. To investigate the effectiveness of our domain adaptation method, we implement different variants of the modality-specific domain adaptation (MSDA) in Table 3. Here, "w/o DA" denotes that we do not design domain adaptation strategy across two datasets; "w/ normal DA" denotes that we follow previous works to directly close the two domain distribution. It shows that both two variants achieve worse performance than our domain strategy "w/ our DA". It demonstrates that an appropriate intermediate domain is able to better bridge

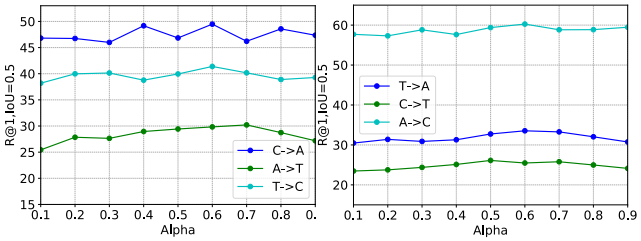

Figure 3: Sensitivity of the hyper-parameter α.

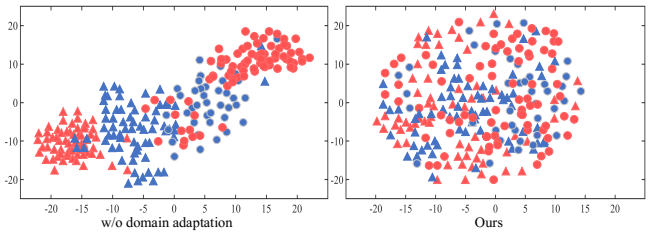

Figure 4: The t-SNE visualizations of "w/o domain adaptation" and our UDA approach. Blue/red denotes source/target domain, while circles/triangles denote videos/queries.

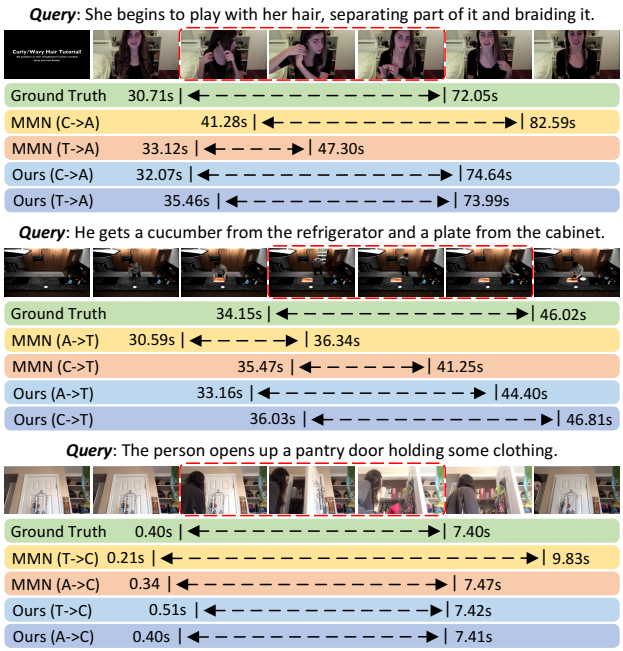

Figure 5: Qualitative results on three datasets.

the two domains to smoothly and accurately reduce the gap.

Effect of the mutual information maximization. To explore the effectiveness of the mutual information maximization (MIM) module, we comprehensively investigate several alignment mechanisms and show the results in Table 3. Specifically, "w/o MIM" denotes that we do not supervise the video-query alignment in target domain. "w/ test" denotes that we directly select the unique text with the highest similarity for each target video for contrastive learning in MIM. Similarly, "w/ video" denotes that we select the unique video with the highest similarity for each target text. we find that both 'w/ test" and "w/ video" brings large improvement to the baseline since they self-supervise the feature learning in target domain. However, it shows that our "w/ mutual" performs the best, demonstrating that the MIM is superior to aligning from only one modality stream.

Sensitivity of hyper-parameters. As shown in Figure 3, we conduct experiments under the setting of all UDA tasks, and present the ablation study on the hyper-parameters α. From this figure, we can find that, within a wide range of α in $[0.1, 0.9]$, the performance only varies in a small range, indicating the robustness to different choices of α. Therefore, we choose $\alpha = 0.6$ in our all experiments.

Visualization

Feature visualization. To investigate the domain distributions of the modalities in each domain, we randomly choose 50 video-query pairs in both source and target domains respectively, and show the t-SNE (Van der Maaten and Hinton 2008) visualizations of "w/o domain adaptation" variant and our UDA variant in Figure 4. We can find that there is a large distribution gap of video/query between the source and target domains of "w/o domain adaptation" variant. Different from it, our proposed UDA method effectively reduces the domain shift by proposing the modality-specific domain adaptation module, thus our learned feature distributions of two domains are mixed up.

Localization visualization. We further provide the localization visualizations of three datasets in Figure 5. It shows that the previous SOTA TSL method MMN fails to localize the accurate query-related segment in the UDA setting. This is because their trained model is data-dependent and suffers from both the cross-domain gaps and the unannotated target correspondences. Instead, our proposed UDA-TSL method reduces such domain shifts in a smooth way, and learns the potential multi-modal correspondences in the target domain, leading to better localization results.

Conclusion

In this paper, we focus on unsupervised domain adaptative temporal sentence localization (UDA-TSL) task. To address it, we propose a new yet effective UDA-TSL baseline for this special task. Specifically, we first propose a modality-specific domain adaptation module to simultaneously generate discriminative multi-modal features and alleviate cross-data domain shifts. Then, we learn the potential correspondence in the unannotated target data by developing a mutual information maximization module to progressively mine truly aligned target pairs and ensure the discriminability of target features. Experiments on three benchmarks demonstrate the effectiveness of our method.

Acknowledgments

This work was supported by the Pioneer and Leading Goose RD Program of Zhejiang (No. 2023C01212), Young Elite Scientists Sponsorship Program by CAST (No. 2022QNRC001).

References

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *ICCV*.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Chao, W.-L.; Hu, H.; and Sha, F. 2018. Cross-dataset adaptation for visual question answering. In *CVPR*.

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally grounding natural sentence in video. In *EMNLP*.

Chen, L.; Lu, C.; Tang, S.; Xiao, J.; Zhang, D.; Tan, C.; and Li, X. 2020a. Rethinking the Bottom-Up Framework for Query-based Video Localization. In *AAAI*.

Chen, M.-H.; Kira, Z.; AlRegib, G.; Yoo, J.; Chen, R.; and Zheng, J. 2019. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*.

Chen, M.-H.; Li, B.; Bao, Y.; AlRegib, G.; and Kira, Z. 2020b. Action segmentation with joint self-supervised temporal domain adaptation. In *CVPR*.

Chen, T.-H.; Liao, Y.-H.; Chuang, C.-Y.; Hsu, W.-T.; Fu, J.; and Sun, M. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *ICCV*.

Chen, Z.; Ma, L.; Luo, W.; Tang, P.; and Wong, K.-Y. K. 2020c. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*.

Choi, J.; Sharma, G.; Schulter, S.; and Huang, J.-B. 2020. Shuffle and attend: Video domain adaptation. In *ECCV*.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NeurIPS*.

Damodaran, B. B.; Kellenberger, B.; Flamary, R.; Tuia, D.; and Courty, N. 2018. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022a. Partially Relevant Video Retrieval. In *ACM MM*.

Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; and Wang, M. 2022b. Dual encoding for video retrieval by text. *TPAMI*.

Fang, X.; and Hu, Y. 2020. Double self-weighted multiview clustering via adaptive view fusion. *arXiv preprint arXiv:2011.10396*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. 2021a. Animc: A soft approach for autoweighted noisy and incomplete multiview clustering. *TAI*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2020. V ́3 H: View variation and view heredity for incomplete multiview clustering. *TAI*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2021b. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *TETCI*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Cheng, Y.; Tang, K.; and Zou, K. 2023a. Annotations Are Not All You Need: A Cross-modal Knowledge Transfer Network for Unsupervised Temporal Sentence Grounding. In *Findings of EMNLP*.

Fang, X.; Liu, D.; Zhou, P.; and Hu, Y. 2022. Multi-Modal Cross-Domain Alignment Network for Video Moment Retrieval. *TMM*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023b. You Can Ground Earlier than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos. In *CVPR*.

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2023c. Hierarchical Local-Global Transformer for Temporal Sentence Grounding. *TMM*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR*.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*.

Gopalan, R.; Li, R.; and Chellappa, R. 2013. Unsupervised adaptation across domain shifts by generating intermediate data representations. *TPAMI*.

Hao, X.; Zhang, W.; Wu, D.; Zhu, F.; and Li, B. 2023. Dual Alignment Unsupervised Domain Adaptation for Video-Text Retrieval. In *CVPR*.

Hosseini-Asl, E.; Zhou, Y.; Xiong, C.; and Socher, R. 2018. Augmented cyclic adversarial learning for low resource domain adaptation. *arXiv preprint arXiv:1807.00374*.

Huang, X.; and Peng, Y. 2018. Deep cross-media knowledge transfer. In *CVPR*.

Huang, X.; Peng, Y.; and Yuan, M. 2018. MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval. *TCYB*.

Lin, Z.; Zhao, Z.; Zhang, Z.; Wang, Q.; and Liu, H. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *AAAI*.

Liu, D.; Fang, X.; Hu, W.; and Zhou, P. 2023a. Exploring Optical-Flow-Guided Motion and Detection-Based Appearance for Temporal Sentence Grounding. *TMM*.

Liu, D.; Fang, X.; Zhou, P.; Di, X.; Lu, W.; and Cheng, Y. 2023b. Hypotheses tree building for one-shot temporal sentence localization. In *AAAI*.

Liu, D.; and Hu, W. 2022. Skimming, Locating, then Perusing: A Human-Like Framework for Natural Language Video Localization. In *ACM MM*.

Liu, D.; Qu, X.; Di, X.; Cheng, Y.; Xu, Z. X.; and Zhou, P. 2022a. Memory-Guided Semantic Learning Network for Temporal Sentence Grounding. In *AAAI*.

Liu, D.; Qu, X.; Dong, J.; Nan, G.; Zhou, P.; Xu, Z.; Chen, L.; Yan, H.; and Cheng, Y. 2023c. Filling the Information Gap between Video and Query for Language-Driven Moment Retrieval. In *ACM MM*.

Liu, D.; Qu, X.; Dong, J.; and Zhou, P. 2021a. Adaptive Proposal Generation Network for Temporal Sentence Localization in Videos. In *EMNLP*.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021b. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding. In *CVPR*.

Liu, D.; Qu, X.; Dong, J.; Zhou*, P.; Xu, Z.; Wang, H.; Di, X.; Lu, W.; and Cheng, Y. 2023d. Transform-Equivariant Consistency Learning for Temporal Sentence Grounding. *TOMM*.

Liu, D.; Qu, X.; and Hu, W. 2022. Reducing the Vision and Language Bias for Temporal Sentence Grounding. In *ACM MM*.

Liu, D.; Qu, X.; Liu, X.-Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020. Jointly Cross-and Self-Modal Graph Attention Network for Query-Based Moment Localization. In *ACM MM*.

Liu, D.; Qu, X.; Wang, Y.; Di, X.; Zou, K.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022b. Unsupervised Temporal Video Grounding with Deep Semantic Clustering. In *AAAI*.

Liu, D.; Qu, X.; and Zhou, P. 2021. Progressively Guide to Attend: An Iterative Alignment Framework for Temporal Sentence Grounding. In *EMNLP*.

Liu, D.; Qu, X.; Zhou, P.; and Liu, Y. 2022c. Exploring Motion and Appearance Information for Temporal Sentence Grounding. In *AAAI*.

Liu, D.; Zhou, P.; Xu, Z.; Wang, H.; and Li, R. 2022d. Few-Shot Temporal Sentence Grounding via Memory-Guided Semantic Learning. *TCSVT*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *ICML*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, 2208–2217. PMLR.

Munro, J.; and Damen, D. 2020. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*.

Na, J.; Jung, H.; Chang, H. J.; and Hwang, W. 2021. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *CVPR*.

Nan, G.; Qiao, R.; Xiao, Y.; Liu, J.; Leng, S.; Zhang, H.; and Lu, W. 2021. Interventional Video Grounding with Dual Contrastive Learning. In *CVPR*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Qu, X.; Tang, P.; Zou, Z.; Cheng, Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *ACM MM*.

Qu, X.; Zou, Z.; Cheng, Y.; Yang, Y.; and Zhou, P. 2019. Adversarial category alignment network for cross-domain sentiment classification. In *NAACL*.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *TACL*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, Y.; Wang, J.; Ma, L.; Yu, Z.; and Yu, J. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, J.; Ma, L.; and Jiang, W. 2020. Temporally Grounding Language Queries in Videos by Contextual Boundary-aware Prediction. In *AAAI*.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. In *AAAI*.

Yang, M.; Zhao, W.; Xu, W.; Feng, Y.; Zhao, Z.; Chen, X.; and Lei, K. 2018. Multitask learning for cross-domain image captioning. *TMM*.

Yuan, Y.; Ma, L.; Wang, J.; Liu, W.; and Zhu, W. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In *NeurIPS*.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In *CVPR*.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2020a. Span-based Localizing Network for Natural Language Video Localization. In *ACL*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020b. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI*.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*.

Zhang, Z.; Zhao, Z.; Lin, Z.; He, X.; et al. 2020c. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. *NeurIPS*.

Zhao, W.; Wu, X.; and Luo, J. 2020. Cross-domain image captioning via cross-modal retrieval and model adaptation. *IEEE Transactions on Image Processing*, 30: 1180–1192.

Zheng, Q.; Dong, J.; Qu, X.; Yang, X.; Wang, Y.; Zhou, P.; Liu, B.; and Wang, X. 2023. Progressive localization networks for language-based moment localization. *TOMM*.

Zhu, J.; Liu, D.; Zhou, P.; Di, X.; Cheng, Y.; Yang, S.; Xu, W.; Xu, Z.; Wan, Y.; Sun, L.; et al. 2023. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. *arXiv preprint arXiv:2301.00514*.