

# OSAN: A One-Stage Alignment Network to Unify Multimodal Alignment and Unsupervised Domain Adaptation

Ye Liu, Lingfeng Qiao, Changchong Lu, Di Yin, Chen Lin, Haoyuan Peng, Bo Ren  
Tencent Youtu Lab

{rafelliu, leafqiao, chonglu, endymecyyin, iwiliamlin, haoyuanpeng, timren}@tencent.com

## Abstract

Extending from unimodal to multimodal is a critical challenge for unsupervised domain adaptation (UDA). Two major problems emerge in unsupervised multimodal domain adaptation: domain adaptation and modality alignment. An intuitive way to handle these two problems is to fulfill these tasks in two separate stages: aligning modalities followed by domain adaptation, or vice versa. However, domains and modalities are not associated in most existing two-stage studies, and the relationship between them is not leveraged which can provide complementary information to each other. In this paper, we unify these two stages into one to align domains and modalities simultaneously. In our model, a tensor-based alignment module (TAL) is presented to explore the relationship between domains and modalities. By this means, domains and modalities can interact sufficiently and guide them to utilize complementary information for better results. Furthermore, to establish a bridge between domains, a dynamic domain generator (DDG) module is proposed to build transitional samples by mixing the shared information of two domains in a self-supervised manner, which helps our model learn a domain-invariant common representation space. Extensive experiments prove that our method can achieve superior performance in two real-world applications. The code will be publicly available.

## 1. Introduction

With explosively emerging multimedia data on the Internet, the field of multimodal analysis achieves more and more attention [10, 13, 18, 19, 43]. Compared to extensive unimodal models in NLP and CV, learning adequate knowledge from multimodal signals is still preliminary but very important. Abundant data plays a key role in different scenarios of multimodal analysis, such as pre-training or downstream multimedia tasks. However, it is prohibitively expensive and time-consuming to obtain large amounts of labeled data. To eliminate this issue, domain adaptation (DA)

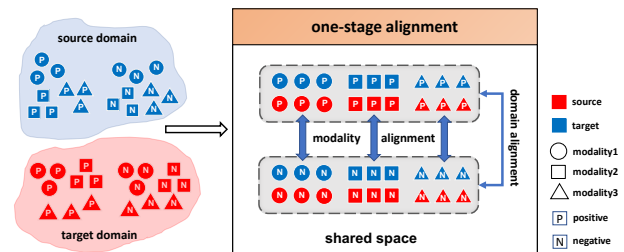


Figure 1. Conception of our one-stage model.

is raised to learn a model from a labeled dataset (source domain) that can be generalized to other related tasks without sufficient labeled data (target domain) [3]. Classical domain adaptation can be classified into different categories: unsupervised domain adaptation (UDA), fully supervised domain adaptation, and semi-supervised domain adaptation [31]. In this paper, we focus on UDA where no samples in target domain are annotated. With this technique, it is not necessary to prepare a customized training dataset for a specific task, but it can perform the task effectively and efficiently.

There are two challenges when applying domain adaptation to multimodal scenarios [17]: (1) how to align the source and target domains and remit domain discrepancy, and (2) how to align multiple modalities and leverage multimodal information. Most existing works address these two problems in two consecutive stages: multimodal alignment followed by domain adaptation [34, 41], or vice versa [14, 44]. However, they solve these two issues separately without considering their relationship: domain and modality can be treated as two views to portray the intrinsic characteristic of multimodal data [8], and the hidden underlying relationship in these two views can provide complementary information to each other. Unimodal domain adaptation methods can not work well in multimodal tasks due to the inability to preserve the relations between modalities at the same time. Through our experiments and analysis, we observe that the two-stage model could not achieve ideal performance. Fig.2 shows the learning curve of two-stage

model during training phase by 800 iterations for the task of multimodal sentiment analysis. It can be found that the learning curve of two-stage model is oscillating and converges slowly, which indicates that two-stage model is probably not a superior solution. To handle these challenges, the objective of multimodal domain adaptation can be defined as: (1) Exploring the relationship between domains and modalities; (2) Finding a common domain-invariant cross-modal representation space to align domains and modalities simultaneously.

Therefore, in this paper, we design a **One-Stage Alignment Network (OSAN)** to unify multimodal alignment and domain adaptation in one stage. Fig.1 shows the conception of our one-stage model. Our method benefits from: (1) The modality and domain are associated and interacted to capture the relationship between domains and modalities, which can provide rich complementary information to each other. (2) Multimodal alignment and domain adaptation are unified in one stage, which allows our model to perform domain adaptation and leverage multimodal information at the same time. In Fig.2, we observe that the learning curve of our method is relatively stable and converges better, which indicates that exploring the relation between modality and domain contributes to our task.

In summary, our contributions are as follows:

(1) To capture the relationship between domain and modality, we propose a one-stage alignment network, called OSAN, to associate domain and modality. In this way, a joint domain-invariant and cross-modal representation space is learned in one stage.

(2) We design a TAL module to bring sufficient interactions between domains and modalities and guide them to utilize complementary information for each other.

(3) To effectively bridge distinct domains, a DDG module is developed to dynamically construct multiple new domains by combining knowledge of source and target domains and exploring intrinsic structure of data distribution.

(4) Extensive experiments on two totally different tasks demonstrate the effectiveness of our method compared to the supervised and strongly UDA methods.

## 2. Related Work

**Unimodal Domain Adaption** The main approaches for domain adaptation or domain alignment can be categorized as discrepancy-based and adversarial-based methods [1, 23, 30]. Discrepancy-based methods design statistics, such as maximum mean discrepancy (MMD) [16], correlation alignment (CORAL) [45] and etc., to measure the difference between two domain distributions. For adversarial-based methods, Ganin et al. [4] used the concept of generative adversarial networks (GAN) to obtain domain-invariant characteristics. Wu et al. [32] proposed a one-stage adaptation framework for nighttime semantic segmentation to per-

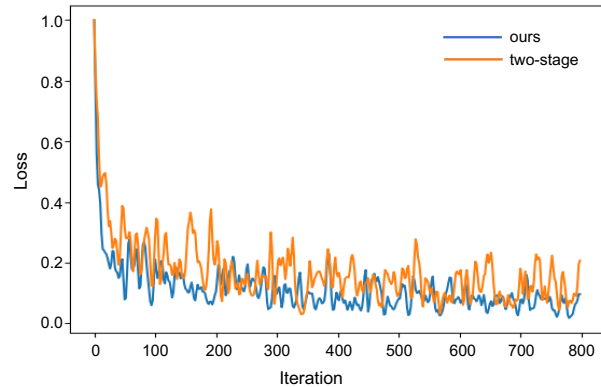


Figure 2. The learning curves of our one-stage method and traditional two-stage method in multimodal sentiment analysis task.

form an end-to-end learning where a separate common pre-processing step was integrated, which is different from our motivation. However, these studies only focus on the elimination of domain shift in unimodal scenarios, which may lose effect when applied in multimodal area.

**Multimodal Domain Adaption** Qi et al. [17] first analyzed the multimodal domain adaptation and raised two problems, i.e., domain adaptation and modality alignment. To alleviate them, hybrid domain constraints and attention-based modality fusion module were introduced to learn domain-invariant fused features. More recently, some studies [14, 34, 41, 42, 44] related to this area have tried to address the multimodal problem by extending the unimodal domain adaptation methods. Xu et al. [34] proposed a supervised multimodal domain adaptation method for VQA to learn joint feature embeddings across different domains and modalities. Munro et al. [14] exploited the correspondence of modalities as a self-supervised alignment approach for UDA in addition to adversarial alignment. Furthermore, Zhou et al. [44] proposed MDMN for early rumor detection, which can combine textual and visual information with two heterogeneous feature extractors. However, all of these studies solved the two problems separately in two stages. Moreover, the relationship between modality and domain is missing, thus they cannot achieve ideal performance.

## 3. Proposed Method

Fig.3 presents an overview of our method that contains four parts: multimodal feature extraction, tensor-based alignment, dynamic domain generator and task-specific heads. First, encoders, as multimodal feature extractor, map source and target multimodal data to different latent spaces. Then, by means of an effective and efficient way of tensor representation, we propose a TAL module to acquire relation information through consecutive interactions between

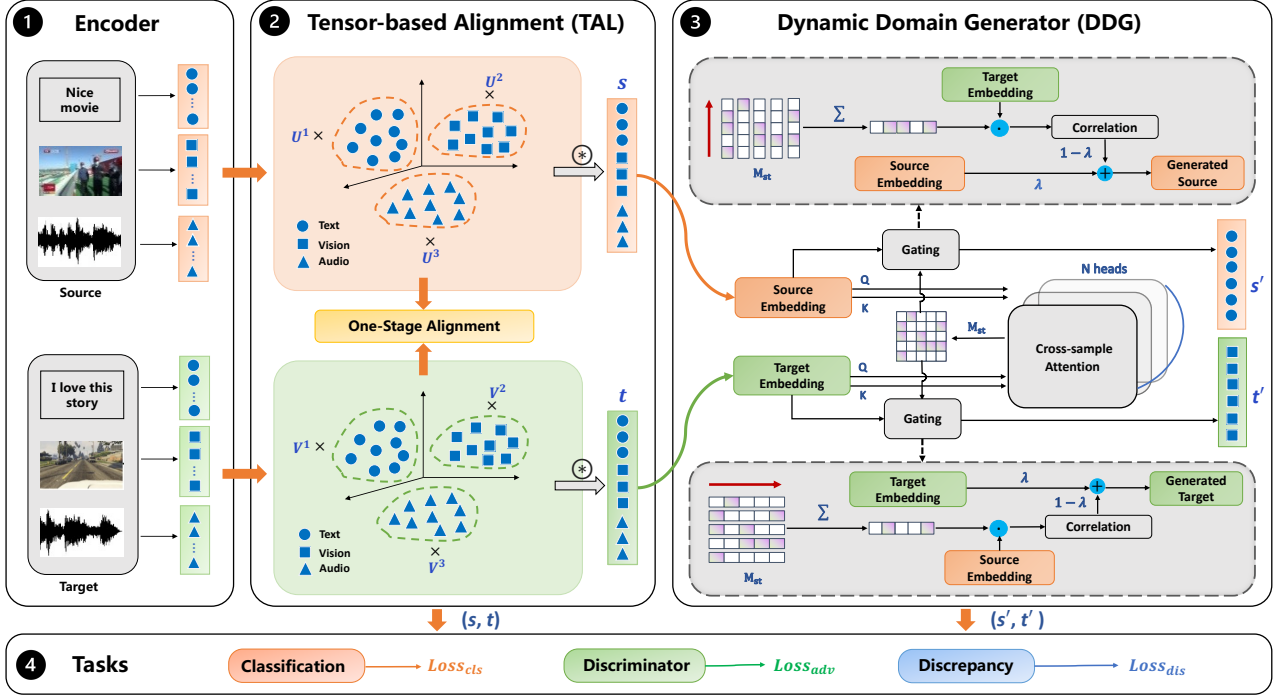


Figure 3. Overview of our method OSAN. We use circle, square and triangle to represent three modalities of text, vision and audio. In the figure,  $\times$  means mode- $d$  product of a tensor and a matrix,  $\odot$  represents element-wise product, and  $\otimes$  means tensor factorization.

modalities and domains. Thus, the modality and domain are aligned simultaneously in one stage. After that, we develop a DDG module to mix information from different domains to create new transitional domains. Finally, our network is split into three branches: category classification, domain adversarial learning and discrepancy elimination.

### 3.1. Tensor-based Alignment

We develop a TAL module to simultaneously address the challenges of multimodal alignment and domain adaptation. We first resort to tensor representation, an effective and natural representation of multimodal data, to develop models capturing inter-modality dynamics [11, 37]. Then, we utilize statistical measurements to establish an interaction between domains and modalities to explore the relation information. In this way, we associate domains and modalities so that they can communicate and align together. Specifically, we perform tensor factorization to fuse multimodal information and maximize statistical measurements to simultaneously eliminate domain gaps.

Specifically, we use  $\mathbf{X}^{I_1 \times \dots \times I_N \times M}$  and  $\mathbf{Y}^{J_1 \times \dots \times J_N \times M}$  to represent source and target dataset with  $N$  modalities and  $M$  samples. For example, a video can be represented by a third-order tensor using a differentiable outer product between visual, textual, and acoustic representations [37]. For simplicity, we first make the training samples zero mean.

To perform multimodal alignment, TAL aims to find pairs of linear transformations  $\mathbf{U}^{(n)}|_{n=1}^N$  and  $\mathbf{V}^{(n)}|_{n=1}^N$  for each modality of source and target domains to project samples of two sets into low dimensional subspaces. During this process, we establish an interaction between domain and modality by maximizing a statistical measurement of covariance given a normalized standard deviation. As reported in [45], we choose covariance instead of the first-order mean statistics used in MK-MMD [5, 21], as it is more powerful and achieves better performance. By this means, domains and modalities can be aligned simultaneously in one stage. Thus the objective function is defined as follows:

$$\begin{aligned}
 & \left\{ \mathbf{U}^{(n)}|_{n=1}^N, \mathbf{V}^{(n)}|_{n=1}^N \right\} \\
 & = \max \left[ \left( \mathbf{X} \prod_{n=1}^N \times_n \mathbf{U}^{(n)T} \right) \otimes \left( \mathbf{Y} \prod_{n=1}^N \times_n \mathbf{V}^{(n)T} \right); \right] \\
 & s.t. \left( \mathbf{X} \prod_{n=1}^N \times_n \mathbf{U}^{(n)T} \right)_{(N+1)}^T \left( \mathbf{X} \prod_{n=1}^N \times_n \mathbf{U}^{(n)T} \right)_{(N+1)} = \mathbf{I} \quad (1) \\
 & \left( \mathbf{Y} \prod_{n=1}^N \times_n \mathbf{V}^{(n)T} \right)_{(N+1)}^T \left( \mathbf{Y} \prod_{n=1}^N \times_n \mathbf{V}^{(n)T} \right)_{(N+1)} = \mathbf{I}
 \end{aligned}$$

The objective function above does not have a closed-form solution, so instead we derive a suboptimal solution following the principle of the alternating projection method [24], where the complicated optimization problem is re-

duced into smaller conditional subproblems that can be solved by simple established methods. Therefore, the objective function is decomposed into  $N$  different subproblems:

$$\begin{aligned} \{\mathbf{U}^{(n)}, \mathbf{V}^{(n)}\} &= \max tr \left\{ \mathbf{U}^{(n)T} \mathbf{C}_{xy}^{(n)} \mathbf{V}^{(n)} \right\} \\ s.t. \quad \mathbf{U}^{(n)T} \mathbf{C}_{xx}^{(n)} \mathbf{U}^{(n)} &= \mathbf{I}, \quad \mathbf{V}^{(n)T} \mathbf{C}_{yy}^{(n)} \mathbf{V}^{(n)} = \mathbf{I} \end{aligned} \quad (2)$$

where, for simplification, we define  $\mathbf{C}_{xy}^{(n)}, \mathbf{C}_{xx}^{(n)}$  as

$$\begin{aligned} \mathbf{C}_{xy}^{(n)} &= \left( \llbracket (\mathbf{X}^{(\bar{n})}) \otimes (\mathbf{Y}^{(\bar{n})}); (\bar{n})(\bar{n}) \rrbracket \right) \\ \mathbf{C}_{xx}^{(n)} &= \left( \llbracket (\mathbf{X}^{(\bar{n})}) \otimes (\mathbf{X}^{(\bar{n})}); (\bar{n})(\bar{n}) \rrbracket \right) \end{aligned} \quad (3)$$

where  $\llbracket \cdot \otimes \cdot; (\bar{n})(\bar{n}) \rrbracket$  means tensor contraction on all indices except the  $n$ -th index on the tensor product [25]. Similarly,  $\mathbf{C}_{yy}^{(n)}$  and  $\mathbf{C}_{yx}^{(n)}$  is same as  $\mathbf{C}_{xx}^{(n)}$  and  $\mathbf{C}_{xy}^{(n)}$  in form. By Eq.2, an interaction between source and target domains for the  $n$ -th modality is established.

### 3.2. Dynamic Domain Generator

In domain adaptation, adversarial learning is often used to align domains by developing a domain discriminator to judge samples from the source or target domains [27, 33]. By this means, it brings a two-class classification task with hard assignments of  $\mathbf{I}$  or  $\mathbf{0}$ . They try to confuse the domain discriminator to learn a shared common representation space between the source and target domains. However, it usually achieves worse performance due to a huge variance between the two domains.

To overcome this problem, several works have been proposed to produce some transitional domains based on Mixup [35, 40]. They mixed images from the source and target domains to generate more soft and mixed samples [26, 33]. Suppose there are one sample  $s$  from source domain and another sample  $t$  from target domain,  $s$  and  $t$  are linearly interpolated [40] to generate a new sample. This strategy can fully utilize the inter-domain information and improve the generalization ability of models. However, it has an obvious shortcoming: the generated sample fuses full information of two domains, which contains redundant information and speciality of domains that may induce undesirable oscillations on training the domain discriminator.

To solve this issue, we propose a DDG module to establish a bridge between domains. Specifically, DDG explicitly captures commonality and abandons the specialty of domains. By highlighting this commonality, we make the domain discriminator focus on commonality rather than full information, which helps our model learn a domain-invariant common representation space.

Given one sample  $s$  from source domain and another sample  $t$  from target domain, DDG is proposed to dynamically generate soft samples  $s'$  and  $t'$  to smoothly bridge the

domain gap. Specifically,  $s'$  represents a variety of  $s$  which maintains information of  $s$  and mix source-relevant commonality from  $t'$ . Similarly,  $t'$  is a variety of  $t$ . As shown in Fig.3, we capture the commonality between  $s$  and  $t$  with the help of the popular and effective self-attention mechanism used in Transformers [29]. We construct a sequence in which the length is 2 by stacking  $s$  and  $t$ . Denote query  $Q_s$  and key  $K_s$  is projected from the source embedding  $s$  respectively, while query  $Q_t$  and key  $K_t$  are from the target embedding  $t$ , respectively. With dot-product between queries and keys, the source-target attention matrix  $M_{st}$  is obtained, which represents the relations between source and target. Afterwards, we design a gating to highlight the commonality between  $s$  and  $t$ , and fuse the commonality to generate  $s'$  and  $t'$  respectively. In detail, to generate  $s'$ , we calculate the feature-relevant distribution  $p$  by summing  $M_{st}$  along the dimension of the query. Larger value in  $p$  means related element in  $t$  is more-relevant to  $s$ , which should be chosen for commonality. As a result, the commonality  $c$  is obtained by  $c = t \odot Norm(p)$ , where  $Norm(\cdot)$  means the normalized operator and  $\odot$  is the element-wise multiplication. Finally,  $s'$  is obtained by fusing commonality  $c$  and information from  $s$ .

$$s' = \lambda s + (1 - \lambda)c \quad (4)$$

Finally, we use multi-head attention to generate various new domains to enrich the sample space. Samples from new domains and raw source and target domains are fed to domain discriminator, by which the domain discriminator is guided by the hard label information and well-designed soft domains. Each sample from these soft domains explores the intrinsic structure of data distribution from raw domains and enriches feature patterns by the interaction of two domains.

### 3.3. Tasks

Our model consists of three tasks: category classification, domain adversarial learning, and domain discrepancy elimination. First, we perform  $K$ -way object classification on the source domain. Second, a domain discriminator is introduced to judge samples from source, target or the created soft domains. Finally, we try to learn transferable features by minimizing domain discrepancy. In summary, we have:

$$L = L_{cls} + \alpha L_{adv} + \beta L_{dis} \quad (5)$$

where  $L_{cls}$  is the category classification loss,  $L_{adv}$  is the adversarial loss, and  $L_{dis}$  is the discrepancy loss.  $\alpha$  and  $\beta$  are weights to balance these three losses. Cross entropy loss is used as classification and adversarial loss, while MK-MMD [5] is employed as discrepancy loss.

## 4. Experiments

We conduct experiments on two totally different tasks to demonstrate the efficacy of our model. One task is popu-

Table 1. Multimodal sentiment analysis results on CMU-MOSEI and CMU-MOSI. †: results come from [7]; ‡: results come from [36]; ◇: results come from [6]; ↓: the lower the better.

Methods	CMU-MOSEI → CMU-MOSI					CMU-MOSI → CMU-MOSEI				
	MAE ↓	Corr	Acc-7	Acc-2	F1	MAE ↓	Corr	Acc-7	Acc-2	F1
<i>Direct Transfer</i>	0.794	0.764	39.5	79.7/81.5	79.5/81.4	0.621	0.685	51.3	79.54/82.14	80.84/81.33
<i>Supervised</i>										
TFN [37] †	0.901	0.698	34.9	-/80.8	-/80.7	0.593	0.700	50.2	-/82.5	-/82.1
ICCN [22] †	0.862	0.714	39.0	-/83.0	-/83.0	0.565	0.713	51.6	-/84.2	-/84.2
MISA [7] ‡	0.804	0.764	-	80.79/82.10	80.77/82.03	0.568	0.724	-	82.59/84.23	82.67/83.97
MAG-BERT [20] ‡	0.731	0.789	-	82.50/84.30	82.60/84.30	0.539	0.753	-	83.80/85.20	83.70/85.10
Self-MM [36] ‡	0.713	0.798	-	84.00/85.98	84.42/85.95	0.530	0.765	-	82.81/85.17	82.53/85.30
MMIM [6] ◇	0.700	0.800	46.65	84.14/86.06	84.00/85.98	0.526	0.772	54.24	82.24/85.97	82.66/85.94
<i>UDA</i>										
DAN [12]	0.777	0.774	39.79	80.03/81.71	79.74/81.49	0.614	0.693	51.6	80.24/81.32	81.36/82.47
ADDA [27]	0.784	0.773	40.14	80.12/82.26	80.13/82.32	0.636	0.707	51.4	80.47/81.59	81.53/82.76
MM-SADA [15]	0.787	0.769	40.52	80.9/82.77	80.68/82.63	0.667	0.684	52.1	80.32/81.44	81.26/81.95
MDMN [44]	0.778	0.774	39.65	81.92/82.01	81.97/82.11	0.602	0.712	52.8	82.24/82.38	82.95/83.26
OSAN(TAL + Mixup)	0.753	0.782	42.64	82.44/83.32	82.14/83.21	0.542	0.757	53.14	82.76/82.88	83.13/83.96
OSAN(TAL + DDG)	<b>0.713</b>	<b>0.801</b>	<b>46.38</b>	<b>83.12/84.58</b>	<b>83.02/84.51</b>	<b>0.532</b>	<b>0.768</b>	<b>53.84</b>	<b>83.41/84.36</b>	<b>83.31/84.47</b>

lar multimodal sentiment analysis, the other is a new task called video text classification.

#### 4.1. Multimodal Sentiment Analysis

Multimodal sentiment analysis is a new dimension of traditional text-based sentiment analysis, which goes beyond the analysis of texts, and includes other modalities data. This task processes data from multiple sources, such as acoustic, visual, and textual information to understand various human emotions.

##### 4.1.1 Experimental Setup

**Datasets and Model Structure** We use two classical datasets, called CMU-MOSEI [39] and CMU-MOSI [38], for the multimodal sentiment analysis task.

**Metrics** We use the metrics that have been presented in [6]: mean absolute error (MAE), which is the average mean absolute difference value between predicted values and truth values. Pearson correlation (Corr) that measures the degree of prediction skew, seven-class classification accuracy (Acc-7) indicating the proportion of predictions that correctly fall into the same interval of seven intervals between -3 and +3 as the corresponding truths, binary classification accuracy (Acc-2) and F1 score for positive/negative and non-negative/negative classification results.

**Baselines** To inspect the relative performance of our model, we compare our model with various baselines in three aspects: **Direct Transfer**, **UDA** and **Supervised**

methods. Direct transfer means that we train a model using source dataset and then predict test samples in target dataset. For UDA, we use model structures of discrepancy-based model DAN [12], adversarial-based model ADDA [27], and some typical studies focused on multimodal area such as MM-SADA [14] and MDMN [44] where multimodal alignment and domain adaptation are performed in two separate stages. Besides direct transfer and UDA methods, we also bring supervised methods for evaluation. For supervised methods, we consider the pure learning-based model TFN [37], feature space manipulation like ICCN [22], MISA [7], more recent and competitive baselines including MAG-BERT [20], Self-MM [36] and MMIM [6].

##### 4.1.2 Experimental Results

**Overall** The overall results are shown in Table 1. In detail, we find that OSAN achieves the best results among all UDA methods on all metrics with large margin. Surprisingly, OSAN even achieves better performance than many supervised methods, such as TFN, MISA and MAG-BERT.

**Ablation Study** To verify the contributions of each component of OSAN, we design a series of ablation experiments. First, we evaluate the effectiveness of the two proposed modules TAL and DDG by eliminating one or both of them from our model, as illustrated in Table 2. We notice a manifest performance degradation after removing TAL or DDG, and the results are even worse when both modules are removed. Furthermore, we conduct additional experiments to explore the contributions of the three losses. In Table 3,

we find that only using classification loss  $L_{cls}$ , our model can achieve competitive performance compared to traditional multimodal DA methods such as MDMN, demonstrating that both modality and domain are well aligned using TAL and DDG. Finally, we adjust some important hyperparameters of DDG and TAL to evaluate their importance. In terms of DDG, we adjust the number of created soft domains by changing multi-head  $H$ . Furthermore, we adjust the fusion coefficient  $\lambda$  to control the fusion ratio between the source and target samples. In Table 4, we observe that the performance improves significantly as the number of domains increases. We also find that the fusion coefficient set to 0.85 can achieve the best result. In terms of TAL, we reduce multimodal features to different combinations of dimensions. We find that the combination of  $384 * 2 * 2$  achieves the best performance, as shown in Table 5.

Table 2. Ablation study of the proposed modules on CMU-MOSI.

Module	w/o DDG	w/o TAL	w/o Both
MAE ↓	+0.039	+0.051	+0.074
Corr	-0.023	-0.021	-0.032
Acc-7	-3.55	-4.65	-5.86
Acc-2	-1.03/-1.30	-1.89/-1.72	-2.22/-1.81
F1	-1.15/-1.52	-1.85/-1.61	-2.34/-1.88

Table 3. Ablation study of three losses on CMU-MOSI.

Loss	$L_{cls}$	$L_{cls} + L_{adv}$	$L_{cls} + L_{dis}$
MAE ↓	+0.018	+0.009	+0.016
Corr	-0.016	-0.012	-0.014
Acc-7	-1.07	-0.82	-0.67
Acc-2	-0.65/-0.95	-0.32/-0.28	-0.67/-0.74
F1	-0.48/-0.69	-0.7/-0.49	-0.31/-0.58

Table 4. Ablation study of DDG hyperparameters on CMU-MOSI.

DDG	$\lambda = 0.85$	$\lambda = 0.9$	$\lambda = 0.85$	$\lambda = 0.9$
	$H = 64$	$H = 64$	$H = 32$	$H = 32$
MAE ↓	0.00	+0.004	0.001	+0.006
Corr	0.00	-0.01	-0.003	-0.016
Acc-7	0.00	-0.590	-0.27	-0.75
Acc-2	0.00/0.00	-0.27/-1.14	-0.16/-1.07	-0.39/-1.29
F1	0.00/0.00	-0.259/-0.15	-0.14/-0.04	-0.39/-0.30

## 4.2. Video Text Classification

Video text classification is a real industrial task for video understanding [9]. With massive videos generated everyday, video textual information extraction is an essential work in many applications. However, numerous useless texts, such as rolling texts and blurred background texts,

Table 5. Ablation study of TAL hyperparameters on CMU-MOSI.

TAL	$t : 384$ $v : 2, a : 2$	$t : 192$ $v : 4, a : 2$	$t : 192$ $v : 2, a : 4$
MAE ↓	0.000	+0.007	+0.006
Corr	0.00	-0.009	-0.006
Acc-7	0.00	-1.17	-0.91
Acc-2	0.00/0.00	-0.77/-1.16	-0.30/-0.91
F1	0.00/0.00	-0.59/-2.02	-0.13/-0.79

have no good effect or even side effects on downstream tasks. To control and filter video content, we propose a new video text classification task to classify video texts into valuable categories.

### 4.2.1 Experimental Setup

**Datasets and Model Structure** We construct a well-defined industrial-grade dataset, called Text-news, which is dedicated to promote video text extraction research for news applications. The extraction of text in videos usually involves two steps: (1) text recognition and (2) text classification. We prepare various videos from news programs for annotation and use a general Optical Character Recognition (OCR) engine to locate and recognize all the texts in the videos. Then, an annotation system is designed for the video text classification task to tag video texts. Three representative categories are defined which are *Caption*, *Subtitle* and *Others*. To verify the effectiveness and generality of our method, we use Text-news as source domain dataset and construct a dataset named Text-show with a few annotated samples as the target domain dataset. The difference between the source and target datasets is that the video program is totally different. The source dataset is constructed from news videos, while the target dataset is constructed from videos of variety shows. We use the standard Precision, Recall, and F1 score for evaluation.

**Baselines** To inspect the relative performance of our model OSAN, we compare it with various baselines in two aspects: **Direct Transfer** and **UDA** methods. Unlike multimodal sentiment analysis, we do not use **Supervised** methods for comparison because the target dataset does not have annotated training data.

### 4.2.2 Experimental Results

**Overall** The results are shown in Table 6. All the metrics of OSAN achieve the best results when compared to other UDA methods and direct transfer, leading to a splendid performance of video text classification task.

Table 6. Video text classification results on Text-show.

Methods	Text-news $\rightarrow$ Text-show		
	Precision	Recall	F1
<i>Direct Transfer</i>	80.2	77.98	79.08
<b>UDA</b>			
DAN [12]	87.44	80.58	83.87
ADDA [27]	91.66	83.66	87.48
MM-SADA [15]	94.07	83.49	88.46
MDMN [44]	93.54	83.69	88.34
OSAN(TAL + Mixup)	94.42	84.79	89.35
OSAN(TAL + DDG)	<b>95.03</b>	<b>86.44</b>	<b>90.53</b>

Table 7. Ablation study of the proposed modules on Text-show.

Module	w/o DDG	w/o TAL	w/o Both
Precision	-0.90	-0.51	-0.96
Recall	-1.24	-2.37	-2.95
F1	-1.09	-1.55	-2.07

Table 8. Ablation study of three losses on Text-show.

Loss	$L_{cls}$	$L_{cls} + L_{adv}$	$L_{cls} + L_{dis}$
Precision	-0.35	-0.17	-0.45
Recall	-0.42	-0.24	-0.14
F1	-0.39	-0.21	-0.28

Table 9. Ablation study of DDG hyperparameters on Text-show.

DDG	$\lambda = 0.9$	$\lambda = 0.95$	$\lambda = 0.9$	$\lambda = 0.95$
	$H = 128$	$H = 128$	$H = 64$	$H = 64$
Precision	0.00	-0.16	-0.75	-0.67
Recall	0.00	-0.32	-1.21	-1.33
F1	0.00	-0.25	-1.01	-1.03

Table 10. Ablation study of TAL hyperparameters on Text-show.

TAL	$t : 56, v : 14$	$t : 14, v : 56$	$t : 28, v : 28$
Precision	0.000	-0.96	+0.24
Recall	0.000	+0.05	-0.59
F1	0.000	-0.41	-0.22

**Ablation Study** Similar to multimodal sentiment analysis, in order to show the benefits of TAL and DDG modules, we perform a series of ablation experiments. First, we eliminate one or both modules from our model. In Table 7, the overall F1 score decreases after removing single module, and the results are even worse when removing both modules, which demonstrates the efficacy of these two modules. Furthermore, we study the impact of three losses of category classification loss, adversarial loss, and discrepancy loss, as shown in Table 8. The results show that all three losses contribute to performance. Moreover, by removing

DA related losses such as adversarial loss and discrepancy loss, it performs slightly worse than keeping all three losses, but still better than traditional multimodal DA methods such as MM-SADA and MDMN. This observation indicates that different domains are well aligned by TAL and DDG. Finally, we tuned some hyperparameters of TAL and DDG to explore their impact. For the DDG hyperparameters, in Table 9, we observe that a smaller fusion coefficient  $\lambda$  and a larger multi-head  $H$  can achieve better results. For the TAL hyperparameters, we reduce multimodal features to different dimensions. In Table 10, the combination  $56 * 14$  for textual and visual features achieves the best result.

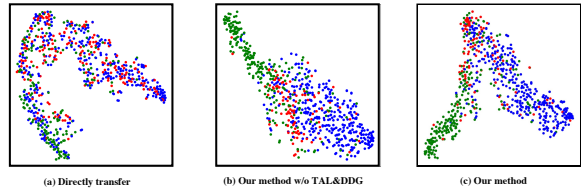


Figure 4. Visualization of the feature distribution of the target domain for multimodal sentiment analysis.

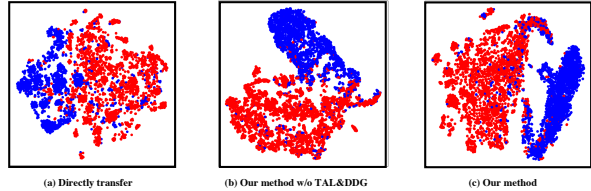


Figure 5. Visualization of the feature distribution of the target domain for video text classification.

## 5. Discussion

### 5.1. Visual Analysis of Feature Distribution

We use t-SNE [28] to visualize the feature distribution of test data in target domain. Fig.5 illustrates the t-SNE visualization of the feature distribution in the video text classification task. We find that the features of different classes are separated more clearly, and the features in the same class are more centralized. A similar phenomenon is observed in Fig.4 of the feature distribution for the multimodal sentiment analysis task. This makes sense because TAL can explore the relationship between domains and modalities and guide them to utilize complementary information. Moreover, DDG can construct new domains that guarantee domain-invariance in a more continuous latent space.

### 5.2. Distribution Discrepancy

To quantify the distribution discrepancy of source and target domains, we use the classical metric of symmetric Jensen-Shannon divergence (JSD) [2]. Fig.7 shows the feature distribution discrepancy based on two baselines and







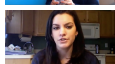
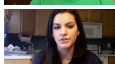


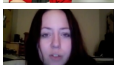
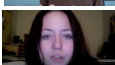
Domain	Image	Text	MDMN Result	Image	Text	OSAN(ours) Result
Source		Beyond that, I actually excelled academically from some of my decisions	gt: 0 prediction: --		Thank you for that Godelieve, we all <b>enjoyed</b> your comments!	gt: 2 prediction: --
Target		um i say go see it coz i i enjoyed it now	gt: 2.0 prediction: -0.37		um i say go see it coz i i <b>enjoyed</b> it now	gt: 2.0 prediction: 1.85
Source		hey were all randomly chosen ad we're going to have my lovely assistant here choose from among these three people	gt: 0 prediction: --		This story demonstrates how <b>ridiculous</b> and petty fighting, or war...	gt: -2 prediction: --
Target		and um i liked you know catherine hardwicke is like you know kind of ridiculous	gt: -0.6 prediction: 0.42		and um i liked you know catherine hardwicke is like you know kind of <b>ridiculous</b>	gt: -0.6 prediction: -0.56
Source		well if they had kept their mouth shut, that movie would have been fabulous.	gt: -1 prediction: --		Learn a language if you can, because that will make your life more <b>interesting</b>	gt: 2.0 prediction: --
Target		and it was kind of interesting	gt: 1.2 prediction: -1.94		and it was kind of <b>interesting</b>	gt: 1.2 prediction: 1.53

Figure 6. Some samples analysis of one baseline MDMN and our model OSAN for UDA task CMU-MOSEI  $\rightarrow$  CMU-MOSI. Columns show information of one representative frame extracted from video clip, text message and the prediction result. Rows give information of target samples and closest source samples.

our model on two tasks. Two baselines are Direct Transfer and our model without the two proposed modules TAL and DDG. We observe that JSD of our model is smaller than JSD of both baselines. This phenomenon implies that our model can reduce the domain gap more effectively than the baselines.

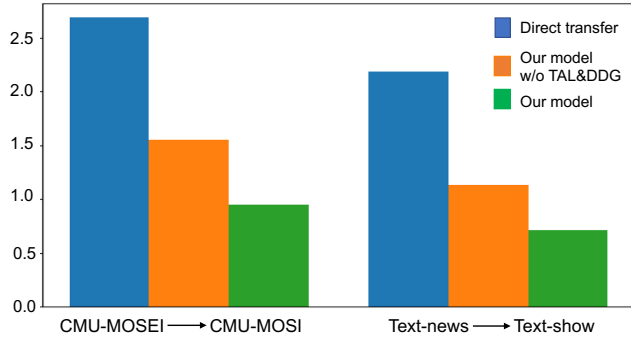


Figure 7. Distribution discrepancy of different methods on two different tasks CMU-MOSEI  $\rightarrow$  CMU-MOSI and Text-news  $\rightarrow$  Text-show.

### 5.3. Case Studies

To deeply explore the advantages of our model over other baselines, we give some predictions of a two-stage baseline MDMN and our model OSAN for the UDA task CMU-MOSEI  $\rightarrow$  CMU-MOSI, as shown in Fig.6. We select some typical samples from the target domain to explore how our model works. For each target sample, we find the closest sample from source domain by calculating cosine similarity between target sample and source sample. We observe that the closest source samples picked by our model OSAN are

more relevant to the target samples. Specifically, take the target sample and source sample in the first row for example, with our model, speakers in both the source frame and target frame are very happy. Moreover, there is one common keyword in what the speakers said: **enjoyed**. Combining these two modalities of image and text, our model correctly classifies this target sample as positive sentiment with a score of 1.85, which is very close to the ground truth score of 2. In contrast, source sample found by the MDMN is not relevant to the target sample. There is no similar semantic information in what they said. The expressions in the frames are dissimilar. The phenomenon in these cases implies that our model can learn a common latent space in which the source domain and target domain are well aligned.

## 6. Conclusion

To address the problems of domain adaptation and modality alignment in multimodal domain adaptation, we propose a novel one-stage alignment network to unify these two problems in one stage. In our model, the TAL module is proposed to explore the relationship between domains and modalities, which can provide complementary information to each other. Moreover, the DDG module is designed to generate new soft samples that mix information from both the source and target domains, guiding our model in a more continuous space. Extensive experiments on two multimodal tasks with different extents of domain shift demonstrate the excellent performance of our model.



## References

- [1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. [2](#)
- [2] Bent Fuglede and Flemming Topsøe. Jensen-shannon divergence and hilbert space embedding. In *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004. [7](#)
- [3] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. [1](#)
- [4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. [2](#)
- [5] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213. Citeseer, 2012. [3](#), [4](#)
- [6] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, 2021. [5](#)
- [7] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020. [5](#)
- [8] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021. [1](#)
- [9] Ye Liu, Changchong Lu, Chen Lin, Di Yin, and Bo Ren. Contrastive graph multimodal model for text classification in videos. *arXiv preprint arXiv:2206.02343*, 2022. [6](#)
- [10] Ye Liu, Lingfeng Qiao, Di Yin, Zhuoxuan Jiang, Xinghua Jiang, Deqiang Jiang, and Bo Ren. Os-msl: One stage multimodal sequential link framework for scene segmentation and classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6269–6277, 2022. [1](#)
- [11] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018. [3](#)
- [12] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. [5](#), [7](#)
- [13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. [1](#)
- [14] 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 122–132, 2020. [1](#), [2](#), [5](#)
- [15] 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 (CVPR)*, June 2020. [5](#), [7](#)
- [16] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010. [2](#)
- [17] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 429–437, New York, NY, USA, 2018. Association for Computing Machinery. [1](#), [2](#)
- [18] Lingfeng Qiao, Chen Wu, Ye Liu, Haoyuan Peng, Di Yin, and Bo Ren. Grafting pre-trained models for multimodal headline generation. *arXiv preprint arXiv:2211.07210*, 2022. [1](#)
- [19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. [1](#)
- [20] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pre-trained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access, 2020. [5](#)
- [21] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. [3](#)
- [22] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999, 2020. [5](#)
- [23] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5940–5947, 2020. [2](#)
- [24] Dacheng Tao, Xuelong Li, Xindong Wu, and Steve Maybank. Tensor rank one discriminant analysis—a convergent method for discriminative multilinear subspace selection. *Neurocomputing*, 71(10):1866–1882, 2008. [3](#)
- [25] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1700–1715, 2007. [4](#)
- [26] Wilhelm Truheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF*

- Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 4
- [27] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 4, 5, 7
- [28] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7
- [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 4
- [30] Wei Wang, Haojie Li, Zhengming Ding, Feiping Nie, Junyang Chen, Xiao Dong, and Zhihui Wang. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2
- [31] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020. 1
- [32] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021. 2
- [33] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6502–6509, 2020. 4
- [34] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multimodal domain adaptation. *arXiv preprint arXiv:1911.04058*, 2019. 1, 2
- [35] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 4
- [36] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *arXiv preprint arXiv:2102.04830*, 2021. 5
- [37] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 3, 5
- [38] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 5
- [39] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018. 5
- [40] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 4
- [41] Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020. 1, 2
- [42] Weichen Zhang, Dong Xu, Jing Zhang, and Wanli Ouyang. Progressive modality cooperation for multi-modality domain adaptation. *IEEE Transactions on Image Processing*, 30:3293–3306, 2021. 2
- [43] Yiyi Zhang, Li Niu, Ziqi Pan, Meichao Luo, Jianfu Zhang, Dawei Cheng, and Liqing Zhang. Exploiting motion information from unlabeled videos for static image action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12918–12925, 2020. 1
- [44] Honghao Zhou, Tinghuai Ma, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. Mdmn: Multi-task and domain adaptation based multi-modal network for early rumor detection. *Expert Systems with Applications*, 195:116517, 2022. 1, 2, 5, 7
- [45] Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. Deep unsupervised convolutional domain adaptation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 261–269, 2017. 2, 3