10-2022

Mix-DANN and dynamic-modal-distillation for video domain adaptation

Yuehao YIN

Bin ZHU
Singapore Management University, binzhu@smu.edu.sg

Jingjing CHEN

Lechao CHENG

Yu-Gang JIANG

Citation

YIN, Yuehao; ZHU, Bin; CHEN, Jingjing; CHENG, Lechao; and JIANG, Yu-Gang. Mix-DANN and dynamic-modal-distillation for video domain adaptation. (2022). *MM '22: Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10-14*. 3224-3233.
Available at: https://ink.library.smu.edu.sg/sis_research/9015

Mix-DANN and Dynamic-Modal-Distillation for Video Domain Adaptation

Yuehao Yin[1,2], Bin Zhu[3], Jingjing Chen[1,2#], Lechao Cheng[4], Yu-Gang Jiang[1,2]
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2] Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]City University of Hong Kong
[4]Zhejiang Lab
yhyin21@m.fudan.edu.cn, {chenjingjing, ygj}@fudan.edu.cn, bin.zhu618@gmail.com, chenglc@zhejianglab.com

ABSTRACT

Video domain adaptation is non-trivial due to video is inherently involved with multi-dimensional and multi-modal information. Existing works mainly adopt adversarial learning and self-supervised tasks to align features. Nevertheless, the explicit interaction between source and target in the temporal dimension, as well as the adaptation between modalities, are unexploited. In this paper, we propose Mix-Domain-Adversarial Neural Network and Dynamic-Modal-Distillation (MD-DMD), a novel multi-modal adversarial learning framework for unsupervised video domain adaptation. Our approach incorporates the temporal information between source and target domains, as well as the diversity of adaptability between modalities. On the one hand, for every single modality, we mix the frames from source and target domains to form mix-samples, then let the adversarial-discriminator predict the mix ratio of a mix-sample to further enhance the ability of the model to capture domain-invariant feature representations. On the other hand, we dynamically estimate the adaptability for different modalities during training, then pick the most adaptable modality as a teacher to guide other modalities by knowledge distillation. As a result, modalities are capable of learning transferable knowledge from each other, which leads to more effective adaptation. Experiments on two video domain adaptation benchmarks demonstrate the superiority of our proposed MD-DMD over state-of-the-art methods.

CCS CONCEPTS

• **Information systems → Multimedia streaming**.

KEYWORDS

Dynamic-Modal-Distillation, Video Domain Adaptation, Adversarial Learning

# indicates corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548313

ACM Reference Format:
Yuehao Yin[1,2], Bin Zhu[3], Jingjing Chen[1,2#], Lechao Cheng[4], Yu-Gang Jiang[1,2]. 2022. Mix-DANN and Dynamic-Modal-Distillation for Video Domain Adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548313

1 INTRODUCTION

Unsupervised domain adaptation (UDA) [44] has drawn significant research attention in recent years, which aims to transfer a model trained on a labelled source domain to an unlabelled target domain with different distribution. UDA is of great value by alleviating the demand to acquire time-consuming and expensive labelled data to train notoriously data-hungry deep neural networks. Existing works [3, 7, 17, 30, 35, 43, 51] have made great progress in image-based domain adaptation. Nevertheless, it is still not sufficiently explored in video-based domain adaptation. Unlike the image, video is naturally multi-modal (*e.g.*, RGB and Optical Flow) and multi-dimensional (*i.e.*, spatial and temporal information), making video domain adaptation much more challenging.

The dominated video domain adaptation methods are mainly based on adversarial learning (AL) [6, 20, 29, 32] due to the simplicity and excellent performance of AL. The AL-based methods add adversarial domain discriminators to distinguish the extracted features from which domain, and the feature extraction network is forced to learn domain-invariant feature representation by adversarially trained with discriminator. AL methods are easy to implement by adding a few standard layers with a gradient reversal layer as a plug-and-play unit [12], which has been widely used for image-based tasks. However, there are very few works to explore how to improve adversarial learning in video domain adaptation. Existing AL-based video domain adaptation methods naively match sample-level feature distributions, which do not make full use of the temporal information between source and target domains. Besides, self-supervised learning (SSL) for feature alignment in cross-domain has been studied extensively in recent years, and significant progress has been made [8, 23, 28, 34, 37]. The SSL-based methods design various self-supervised sub-tasks to enhance the robustness of feature extraction and the alignment between different modalities, then assign pseudo labels to the target data to better align domain-invariant features. Contrastive learning is widely used in self-supervised sub-tasks, which include complicated feature space projections and similarity calculations. Moreover, the performance on the target domain heavily depends on the reliability of the assignment of pseudo labels.

In the literature, it is a common practice to make use of the correlation between different modalities to enhance the domain adaptation performance, as video is related to multiple modalities, for instance, [23, 28, 29, 37]. In addition, it is shown that domain shift in different modalities varies, resulting in the performance of the sub-model for each modality being distinct [28, 29]. [28] makes use of the modal-diverseness through Asynchronous Learning. Its superior performance indicates the potential of exploring the modalities' differences for domain adaptation. However, most existing multi-modal domain adaptation methods [23, 29, 37] treat different modalities equally. How to utilize the difference in adaptability among the modalities is yet to be explored.

This paper addresses the aforementioned limitations for video domain adaptation by combining the multi-dimensional source and target data with adversarial learning and further making use of the modal diversities in videos. The main idea of our approach is shown in Figure 1. Specifically, we propose Mix-Domain-Adversarial Neural Network (Mix-DANN), which extends DANN [12] with mix-samples. For each modality, we firstly mix the source and target samples with a specified ratio along the temporal dimension to produce mix-samples, which are also regarded as inputs for the model like the vanilla-samples (*i.e.*, original samples). Unlike DANN [12], the adversarial domain discriminator not only needs to distinguish which domain the feature comes from when the input is a vanilla-sample but also should identify the mix ratio between source and target domains when the input is a mix-sample. A non-adversarial mix-sample classifier is also added as a self-supervised auxiliary task. As a result, the ability to capture domain-invariant feature representation is further enhanced. Additionally, Dynamic-Modal-Distillation (DMD) is proposed by taking advantage of the diversities within different modalities. Knowledge distillation [18] is employed to transfer the knowledge from the most adaptable modality to other modalities during training. We dynamically add the loss from the mix-sample discriminator and the adversarial losses as Teacher Score to measure the degree of adaptation for each modality. A higher Teacher Score indicates the model is more confusing to get the domain-relevant information from the features, *i.e.*, the features are more domain-invariant and adaptable. After obtaining the adaptability to the target domain of each modality, the most adaptable modality is set to play the role of the teacher to guide the other less adaptable modalities. In the whole training process, modalities guide and reinforce each other alternatively. Finally, the multi-modal model achieves stronger adaptability in the target domain.

We examine the performance of MD-DMD on two video domain adaptation benchmarks, EPIC-Kitchens [9, 10] and UCF-HMDB [24, 38] datasets. Experimental results demonstrate that Mix-DANN improves accuracy in every single modality, and Dynamic-Modal-Distillation can boost the multi-modal model's adaptability. By combining these two components, MD-DMD outperforms state-of-the-art methods for unsupervised video domain adaptation.

The main contributions of this paper are as follows:

- A novel Mix-DANN method is proposed to leverage adversarial learning in video tasks in a single modality, which makes full use of the temporal dimension of video inputs to enhance adversarial strength and let the single-modal model learns more domain-invariant feature representations.
- We propose a novel Dynamic-Modal-Distillation method that dynamically measures the adaptability of each modality during training, and then the adaptability score enables modalities to teach each other domain adaptable knowledge by knowledge distillation. Our method is the first attempt to leverage the knowledge distillation technique to solve multi-modal domain adaptation problems.
- Experimental results demonstrate the effectiveness of the proposed components, and the overall MD-DMD achieves state-of-the-art performance in unsupervised action recognition domain adaptation benchmark datasets.

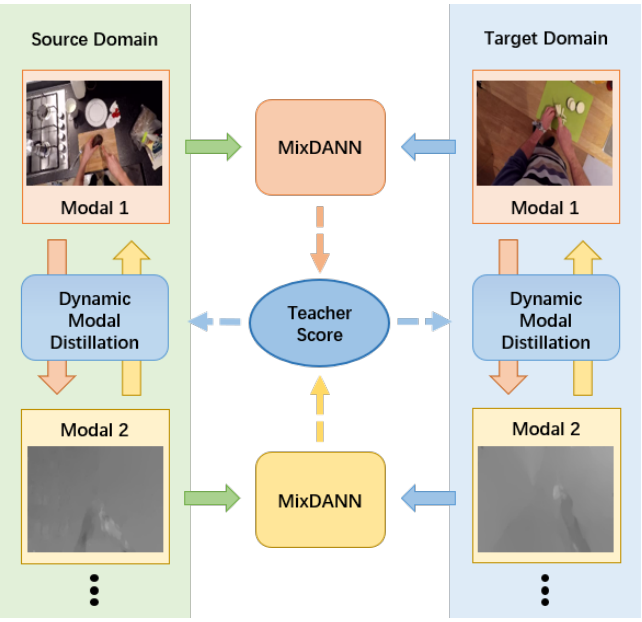

Figure 1: The main idea of our proposed approach. The MixDANN is used to align feature distribution from the source and target domains within a modality and then produces a teacher score to measure the modality's adaptability. Dynamic-Modal-Distillation is used to dynamically transfer knowledge from the most adaptive modality to other less adaptive modalities based on teacher scores.

2 RELATED WORK

The proposed MD-DMD is closely relevant to research areas, including supervised video classification, unsupervised domain adaptation, video domain adaptation, and knowledge distillation.

2.1 Supervised Action Recognition

Action recognition is the basic problem in video understanding, which can be roughly divided into two categories: 2D-CNN [22] or 3D-CNN [21] based methods. Besides RGB frames, Optical Flow is commonly used in action recognition models. Two-stream networks [36] is an effective 2D-CNN based method that provides a basic architecture for multi-modal models. C3D [41] is a milestone

for 3D-CNNs. I3D [5] extends the two-stream networks by inflating the 2D convolution layers to 3D and initializing the corresponding 3D layers with pre-trained 2D CNNs. Besides, several works are proposed to further incorporate the temporal relations between frames, *e.g.*, Non-Local [46], SlowFast [11], Group Contextualization [16]. And a lot of works focus on reducing the high computational cost of 3D convolutions [33, 42, 47–49]. In addition, compared with CNN models, transformers recently show great potential and promising results in action recognition [1, 27, 45, 50]. However, the performance of these models deteriorates significantly when directly applied to domains with different distribution. Our focus is to improve the performance of the action recognition models without extra labels in the target domain.

2.2 Unsupervised Domain Adaptation

Typical unsupervised domain adaptation (UDA) approaches can be summarized into three categories [44]: discrepancy-based, adversarial-based, and reconstruction-based. The discrepancy-based methods work by aligning the statistical distribution shift between the source and target domains, *e.g.*, MMD [13], CORAL [39], KL [52] and BN [19]. The adversarial-based methods use a domain discriminator that distinguishes whether the extracted features are from the source or target domain to encourage domain confusion through an adversarial objective, *e.g.*, DANN [12] and ADDA [43]. The reconstruction-based methods reconstruct the features of target samples, which is helpful for improving the performance on the target domain, *e.g.*, TLDA [52] and DRCN [14]. More recently, several works have designed self-supervised auxiliary tasks to help learn transferable features for DA, *e.g.*, JiGen [4] and [40].

2.3 Video Domain Adaptation

Most UDA methods are proposed mainly for image tasks, while video domain adaptation received less attention until recent years. Several works attempt to extend the UDA methods from image to video case. Adversarial-based methods are widely used due to efficiency and simplicity. DAAA [20] and MM-SADA [29] align the features from different domains by adding a domain-discriminator as DANN [12]. However, they both trivially match the segment-level feature distribution without making use of the temporal dimension in adversarial training. TA3N [6] and TCoN [32] combine the adversarial framework [12] with different attention mechanisms, but their works are not proposed for multi-modal architectures, which are commonly used in video tasks. Besides adversarial-based methods, self-supervised methods, especially contrastive learning, for video DA gains significant attention in recent years. A dizzying variety of self-supervised auxiliary tasks are designed. SAVA [8] uses the clip order prediction as an auxiliary task, which is proposed for RGB single modality. CoMix [34] utilizes temporal contrastive learning by minimizing or maximizing the similarity of different videos or the same video played at different speeds and using background mixing to leverage action semantics shared across both domains. The CoMix is also proposed for RGB single modality, and prior background subtraction techniques are required. STCDA [37] proposes spatio-temporal contrastive learning and video-based contrastive alignment to establish the cross-modal domain alignment, and a cluster algorithm was used to assign pseudo-labels to target data. CrossModal [23] simultaneously uses cross-modal contrastive learning to align cross-modal representations from the same video and cross-domain contrastive learning to align representations between the source and target domains in each modality, pseudo-labels are assigned by setting a certain threshold. The multi-modal models mentioned above are all treated each modality equally. However, the degrees of domain shift and the adaptability of different modalities are usually diverse (Observed in the paper of DLMM [28]). DLMM makes use of the diversity between multiple modalities by proposing a novel Prototype based Reliability Measurement to estimate the reliability of the recognition results to assign pseudo-labels to target data and an Asynchronous Curriculum Learning strategy that chooses the pseudo-labelled target samples from easy to hard to train the sub-models. In contrast, our work aligns cross-domain feature distributions by a novel adversarial based method MixDANN and makes use of the modalities difference by our knowledge distillation based method Dynamic-Modal-Distillation.

2.4 Knowledge Distillation

Knowledge distillation (KD) [18] is proposed for model compression by transferring knowledge between different neural networks. The bigger complex model is called Teacher Model whose knowledge is transferred to a smaller compact Student Model by minimizing their output probability distribution. Temperature-based softmax is used to compute the probability distribution. A higher temperature makes the distribution softer, which can reveal the similarities between different classes, thus it is beneficial for knowledge transfer. Several works have utilized KD for specific DA tasks, *e.g.*, [2] applies KD by trivially transferring the knowledge of the source model to the adapted model on acoustic tasks, [25] uses the posterior probabilities generated by the source-domain model as pseudo-label for target data to train the target-domain model, [31] adopts KD to semi-supervised DA for segmentation of white matter hyperintensities (WMH) in magnetic resonance imaging (MRI) scans generated by scanners, [15] adapts a complex teacher model to a compact student model by progressively teaching the student about domain-invariant features using KD. Nevertheless, all these DA methods employ KD in a single modality, transferring knowledge from one model to another, while our proposed Dynamic-Modal-Distillation utilizes KD in multi-modal domain adaptation by transferring knowledge between modalities, and we propose Teacher Score to control the direction of distillation.

3 METHOD

An overview of our proposed method MixDANN and Dynamic-Modal-Distillation (MD-DMD) is presented in Figure 3. In a nutshell, the MixDANN is particularly built based on DANN [12] by mixing samples from source and target domains to capture domain-invariant feature representation in every single modality (*e.g.*, RGB and Optical Flow). Moreover, the Dynamic-Modal-Distillation module is proposed to distil knowledge across modalities based on transferability for multi-modal domain adaptation.

To be specific, MixDANN extends the well-known DANN from image to video tasks, which takes mix-samples and vanilla-samples

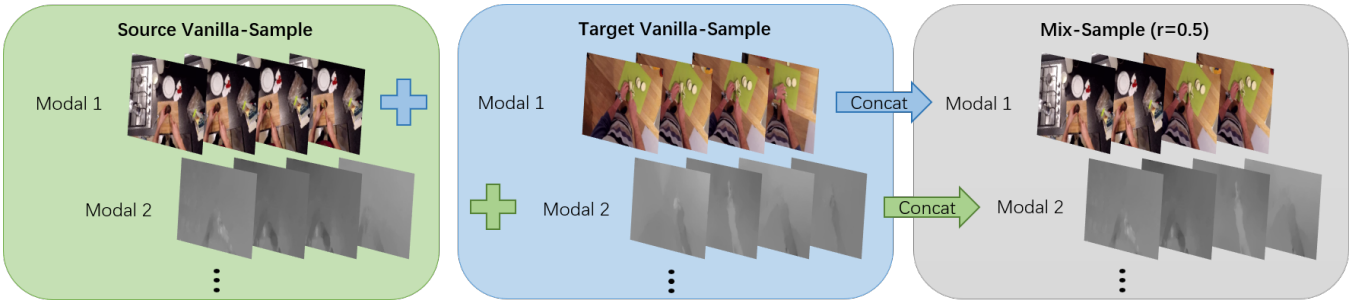

Figure 2: The first two blocks show vanilla-samples from the source (green) and target (blue) domains, respectively. The source vanilla-sample contains label while the target vanilla-sample does not. The block on the right (gray) is one mix-sample generated by mixing vanilla-samples in source and target domains with a specific ratio r (*e.g.*, $r = 50\%$). The frames are concatenated along the temporal dimension for each modality.

from the two domains as inputs. Mix-samples are obtained by mixing the frames from both source and target domains along the temporal dimension. The mix-samples contain information from both source and target domains. The extracted features are more domain-invariant by confusing the adversarial domain-discriminator to predict the mix-ratio. A non-adversarial mix-sample classifier is added on top of the backbone network as a self-supervised auxiliary task. The purpose of MixDANN is two-fold, on the one hand, to enhance the ability of feature learning and, on the other hand, as a basis to measure the adaptability (the degree of adaptation to the target domain) of each modality. In addition, Dynamic-Modal-Distillation is proposed that dynamically measures the adaptability of one modality with the Teacher Score produced by MixDANN during training, and then the adaptability score enables modalities to teach each other domain adaptable knowledge by knowledge distillation. The teacher score controls the direction of Modal-Distillation, where a higher score indicates a modality is more transferable. Based on the value of the Teacher Score, knowledge distillation between modalities is conducted in every iteration during training from a modality with a higher score to others. In the following subsections, we describe each of the proposed components in detail.

3.1 Vanilla-Sample and Mix-Sample

Denote a labelled video set $\boldsymbol{S} = \{X_S, Y_S\}$ as source domain and an unlabelled video set $\boldsymbol{T} = \{X_T\}$ as target domain, where $X =< X^1, X^2, \ldots, X^M >$ indicates multi-modal input sets with M modalities. Y_S refers to the label set of the source domain. Denote $x^m \in \mathbb{R}^{t\times h\times w\times c}$ as an input sample of the $m^{th}(1 \le m \le M)$ modality with t frames. The original samples from the source or target domains, *i.e.*, $X_{vanilla} = \{X_S, X_T\}$ are called **vanilla-samples**.

Labeled source vanilla-samples of the m^{th} modality $\boldsymbol{S}^m$ are used to train the backbone feature extractor $\boldsymbol{F}^m$ and video classifiers $\boldsymbol{C}^m$. Similar with previous works [23, 28, 29], late fusion is adopted for different modalities and cross-entropy loss is employed for supervised classification:

$$\mathcal{L}_C = -\sum_{x\in X_S} y \log \sigma(\sum_{m=1}^{M} C^m(F^m(x^m))), \tag{1}$$

where σ stands for softmax function and $y \in Y_S$ is the one-hot label vector of the source vanilla-sample x.

$$\mathcal{L}_C^m = -\sum_{x\in X_S} y \log \sigma(C^m(F^m(x^m))), \tag{2}$$

Mix-samples (denoted as X_{mix}) are formed by randomly mixing vanilla-samples from source and target domains. Specifically, as shown in Figure 2, we extract $r \times t$ frames from one source vanilla-sample $x_S^m \in X_S$ and $(1-r) \times t$ frames from one target vanilla-sample $x_T^m \in X_T$, where $r \in (0, 1)$ indicates the ratio of source domain of this mix-sample and t is the temporal window of video samples. Then we concatenate the source and target frames along the temporal dimension to form one mix-sample $x_{mix}^m \in \mathbb{R}^{t\times h\times w\times c}$ for the m^{th} modality. Along with vanilla-samples in source and target domains, the mix-samples are fed into the MixDANN with details in the following subsection.

3.2 MixDANN

MixDANN extends DANN [12] to video tasks by introducing mix-samples. In brief, the key idea of DANN is to employ a binary classifier named domain-discriminator D to distinguish whether the inputs from the source or target domain, with a gradient reversal layer to maximize the discriminative loss of feature extractor when minimizing that loss of the discriminator. As shown in Figure 3, different from DANN, we take vanilla-samples $X_{vanilla}$ and mix-samples X_{mix} together as inputs for adversarial training. Each modality has its own discriminator, which can avoid the easier solution of the network focusing only on the less robust modality in classifying the domain [29].

For each modality, when the input is a vanilla-sample the domain-discriminator D^m is to predict the probability which domain the extracted feature comes from, as follows:

$$\mathcal{L}_{D_{vanilla}}^m = -\sum_{x\in\{X_S,X_T\}} y_d \log \sigma(D^m(F^m(x^m))), \tag{3}$$

where the domain label y_d is a binary one-hot vector, $y_d =< 1, 0 >$ for $x \in X_S$ and $y_d =< 0, 1 >$ for $x \in X_T$.

When a mix-sample x_{mix}^m is fed into the MixDANN network, the same domain discriminator D^m is employed to predict the mix-sample's proportion from each domain by its binary output

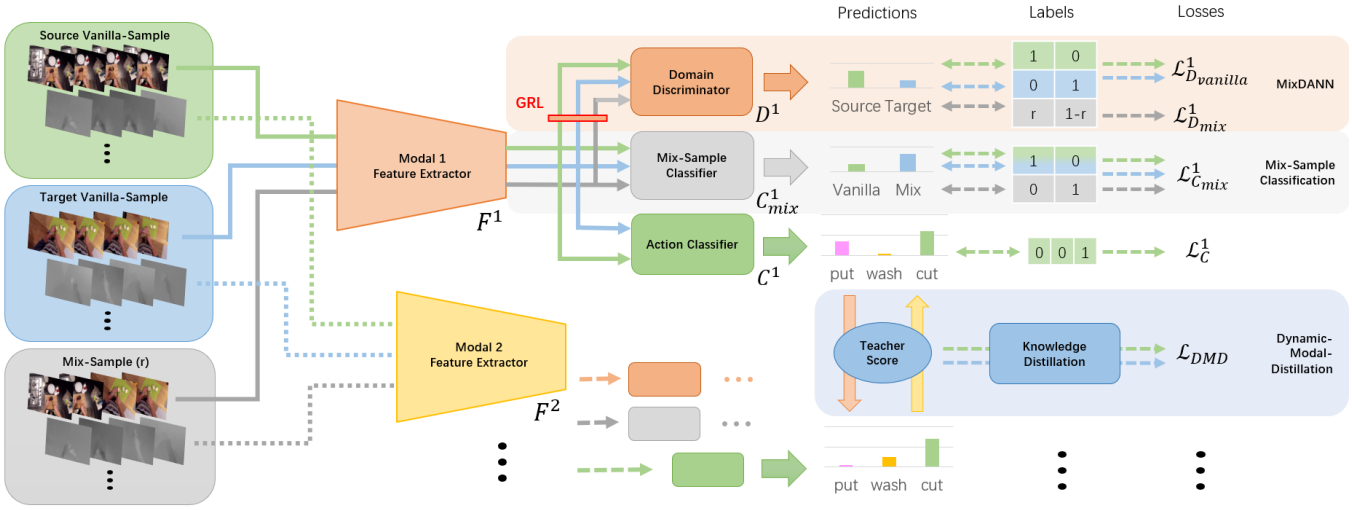

Figure 3: The overview of the proposed Mix-Domain-Adversarial Neural Network and Dynamic-Modal-Distillation (MD-DMD). The three blocks on the left are different types of input samples. We use different colours to indicate the flow direction of different input types. For each modality, all three types of inputs are used to train the adversarial Domain Discriminator and the Mix-Sample Classifier. The Action Classifier is trained by the vanilla-samples from the source domain only. The adversarial losses $\mathcal{L}_{D_{vanilla}}$, $\mathcal{L}_{D_{mix}}$ and the non-adversarial $\mathcal{L}_{C_{mix}}$ are added as Teacher Score $\mathcal{TS}$. The $\mathcal{TS}$ is calculated in every training step for each modality, and the modality with the largest $\mathcal{TS}$ is chosen as the Teacher Modality to teach other modalities by knowledge distillation. All vanilla-samples from both source and target domains can be used in the distillation.

probability distribution. We use y^m_{mix} indicating the groundtruth mix proportion distribution, *i.e.*, $y_{mix} =< r, (1-r) >$. And we use $P(x^m_{mix})$ as the prediction distribution, the mix loss is defined as follows:

$$\mathcal{L}^m_{D_{mix}} = KL(y^m_{mix}||P(x^m_{mix})) = -\sum y^m_{mix} \log(\frac{P(x^m_{mix})}{y^m_{mix}}), \quad \text{where:} \quad P(x^m_{mix}) = \sigma(D^m(F^m(x^m_{mix}))). \tag{4}$$

The MixDANN loss is accumulated by summing both vanilla and mix adversarial losses from all modalities as follows:

$$\mathcal{L}_{MixDANN} = \sum_{m=1}^{M} (\mathcal{L}^m_{D_{vanilla}} + \mathcal{L}^m_{D_{mix}}). \tag{5}$$

3.3 Self-supervised Mix-Sample Classification Task

To enhance the ability of MixDANN to extract domain-invariant features and better measure each modality's adaptability, we introduce a self-supervised mix-sample classification auxiliary task. Specifically, a simple binary classifier C_{mix} is added to each modality after the feature extractor F. The classifier C^m_{mix} of the m^{th} modality is used to predict whether the input sample is vanilla or mixed.

Given a binary one-hot label y_x indicating a sample is vanilla or mixed, the Mix-Sample Classification loss is calculated by cross-entropy:

$$\mathcal{L}^m_{C_{mix}} = -\sum y_x \log \sigma(C^m_{mix}(F^m(x^m))), \tag{6}$$

where $y_x =< 1, 0 >$ for $x \in X_{vanilla}$ and $y_x =< 0, 1 >$ for $x \in X_{mix}$.

The classification losses from all modalities are summed up as the loss of the self-supervised sub-task:

$$\mathcal{L}_{MixCls} = \sum_{m=1}^{M} \mathcal{L}_{C^m_{mix}}. \tag{7}$$

3.4 Teacher Score

If the domain discriminator D is more confusing in determining the domain composition of the sample, it means the extracted feature contains less domain-specific information. In other words, the more domain-invariant the feature is, the more difficult the discriminator to determine and the higher the losses values are. The same for the mix-sample classifier C_{mix}. The higher the values of these losses, the more domain-invariant the features extracted by the feature extractor are, that is, the better the adaptability of this modality. Thus, the value of the losses $\mathcal{L}^m_{D_{vanilla}}$, $\mathcal{L}^m_{D_{mix}}$ and $\mathcal{L}^m_{C_{mix}}$ can indicate the domain-invariant degree of the m^{th} modality. We propose the **Teacher Score** ($\mathcal{TS}$) to measure the adaptability of each modality dynamically during the training process as follows:

$$\mathcal{TS}^m = \mathcal{L}^m_{D_{vanilla}} + \mathcal{L}^m_{D_{mix}} + \mathcal{L}^m_{C_{mix}}. \tag{8}$$

$$\mathcal{TS}^1 \tag{9}$$

$$\mathcal{TS}^2 \tag{10}$$

Based on the teacher score, we can conduct **Dynamic Modal-Distillation** for multi-modal domain adaptation, described in the following subsection.

3.5 Dynamic Modal-Distillation

We propose **Dynamic Modal-Distillation** (DMD) to deal with the problem of modal differences in both the degree of domain shift and the adaptability of sub-modals.

During training, we calculate the teacher score $\mathcal{TS}$ dynamically for each modality as a measure of current adaptability to the target domain for that modality, larger $\mathcal{TS}$ indicating a more adaptive modality. Therefore, we pick the modality with the largest $\mathcal{TS}$ as the **Teacher Modality** to teach the other **Student Modalities**. Knowledge distillation, originally proposed for model compression, is utilized to transfer the currently most domain-invariant knowledge from the Teacher Modality to the less adaptive modalities. Modalities are learning from and teaching each other during training. More specifically, the target of Modal-Distillation loss $\mathcal{L}_{DMD}$ is the soft-distribution of class probabilities predicted by the Teacher Modality, calculated by the softmax function with temperature. The soft-distribution preserves more information about correlations across different classes, which offers better generalizability. Denote $Q(x^t)$ as the output probabilities distribution of Teacher Modality for sample x, $P(x^i)$ as it of the i^{th} Student Modality, and T as the softmax temperature. The DMD loss is defined as:

$$\begin{aligned} \mathcal{L}_{DMD} = \sum_{i=1}^{M-1} KL(Q(x^t)||P(x^i)), \quad x \in \{X_S, X_T\}, i \neq t \\ \text{where: } Q(x^t) = \sigma(C^t(F^t(x^t))/T) \\ \text{and } P(x^i) = \sigma(C^i(F^i(x^i))/T). \end{aligned} \tag{11}$$

3.6 MD-DMD

The overall proposed end-to-end multi-modal domain adaptation framework, MixDANN and Dynamic-Modal-Distillation (MD-DMD), is illustrated in Algorithm 1, and the objective is defined as follows:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{MixDANN} + \mathcal{L}_{MixCls} + \mathcal{L}_{DMD}. \tag{12}$$

Note that the $\mathcal{L}_C$ is computed only on labelled source data. The adversarial $\mathcal{L}_{MixDANN}$ and non-adversarial $\mathcal{L}_{MixCls}$ are computed from both vanilla-samples from the source and target domains and the proposed mix-samples. And the dynamic modal-distillation loss $\mathcal{L}_{DMD}$ is optimised on vanilla source and target data.

4 EXPERIMENTS

In this section, we first introduce the datasets and the implementation details, then we show performance comparisons with the state-of-the-art methods, and finally, we present the ablation study. We adopt Top-1 accuracy (%) as evaluation metric for the experiments by following [23, 28, 29, 34, 37].

4.1 Datasets

We adopt the EPIC-Kitchens [9, 10] and two small-scale UCF[38] and HMDB[24] action recognition datasets to evaluate our proposed MD-DMD framework. Below presents their brief descriptions.

EPIC-Kitchens is a challenging egocentric video dataset which serves as a standard benchmark to test domain adaptation for fine-grained action recognition. We adopt the same domain adaptation settings as previous works [23, 28, 29, 34, 37], focusing on the 8 largest classes in three kitchens D1, D2 and D3, using 2 commonly

Algorithm 1 MixDANN and Dynimaic Modal-Distillation

Input: Labelled source data S and unlabelled target data T

repeat
- Sample batches $\mathcal{B}_s \subset S$ and $\mathcal{B}_t \subset T$
- Generate mix-sample batch $\mathcal{B}_{mix}$ by concatenate vanilla-samples in $\mathcal{B}_s$ and $\mathcal{B}_t$ along the temporal dimension
- Calculate $\mathcal{L}_C$ by $\mathcal{B}_s$
- Calculate $\mathcal{L}_{D_{vanilla}}$ by both $\mathcal{B}_s$ and $\mathcal{B}_t$ for each modality
- Calculate $\mathcal{L}_{D_{mix}}$ by $\mathcal{B}_{mix}$ for each modality
- Calculate $\mathcal{L}_{C_{mix}}$ by $\mathcal{B}_s$, $\mathcal{B}_t$ and $\mathcal{B}_{mix}$ together for each modality
- Calculate the Teacher Score $\mathcal{TS}$ for each modality, pick the modality with the largest $\mathcal{TS}$ as Teacher Modality
- Do knowledge distillation from Teacher Modality to other Student Modalities, Calculate $\mathcal{L}_{DMD}$
- Calculate the over all objective $\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{MixDANN} + \mathcal{L}_{MixCls} + \mathcal{L}_{DMD}$, where $\mathcal{L}_{MixDANN} = \sum_{m=1}^{M}(\mathcal{L}^m_{D_{vanilla}} + \mathcal{L}^m_{D_{mix}})$, $\mathcal{L}_{MixCls} = \sum_{m=1}^{M}\mathcal{L}^m_{C_{mix}}$
- Backpropagate $\mathcal{L}$ then update parameters

until done

employed modalities: RGB and Optical Flow. The number of samples in the three domains is 1978, 3245 and 4871, respectively.

UCF-HMDB are overlapped subsets of the original UCF101 and HMDB51 datasets for action recognition. UCF-HMDB have 12 shared categories.

4.2 Implementation Details

Following previous video domain adaptation works, we use the I3D [5] as our backbone for feature extractor F of each modality. For the Adversarial-Domain-Discriminator D and Mix-Sample Classifier C_{mix}, we use 2 fully-connected layers with hidden layer of 100 dimensions. The size of the input clips is 16 frames with 224×224 pixels. The mix-samples are produced by a random ratio r in our framework. The temperature T of our Dynamic-Modal-Distillation is set to 5 in the comparison results. We train all our models end-to-end. Followed the 'pre-train then adapt' procedure for multi-modal domain adaptation as [28] does, each sub-model of a modality is pre-trained on the labelled source domain independently for the classification task before the proposed approach is employed. The optimizer is stochastic gradient descent (SGD) with momentum of 0.9, and the weight decay is set to $1e-7$. The weights of each loss term are set to 1 equally. On average, training takes 14 hours on 4 RTX 3090 GPUs with batch size 28 and 16000 training steps.

4.3 Comparison with State-of-the-Arts

We compare the proposed MD-DMD with current State-of-the-Art methods (SOTAs) [8, 23, 29, 37] on EPIC-Kitchens and UCF-HMDB datasets in Table 1 and Table 2, respectively.

On the EPIC-Kitchens dataset, the DANN [12] outperforms three commonly used baselines, AdaBN [26], MMD [13] and MCD [35], by employing a domain discriminator with adversarial learning. Compared with DANN, our proposed adversarial based MD-DMD

Table 1: Performance comparison on EPIC-Kitchens dataset.

Method	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	Avg	Gain
Source-only	42.5	44.3	42.0	56.3	41.2	46.5	45.5	-
AdaBN [26]	44.6	47.8	47.0	54.7	40.3	48.8	47.2	+1.7
MMD [13]	43.1	48.3	46.6	55.2	39.2	48.5	46.8	+1.3
MCD [35]	42.1	47.9	46.5	52.7	43.5	51.0	47.3	+1.8
DANN [12]	50.2	47.9	46.5	52.7	43.5	51.0	48.6	+3.1
MM-SADA [29]	46.9	50.2	50.2	53.6	44.7	50.8	50.3	+4.8
CrossModal [23]	49.5	51.5	50.3	**56.3**	46.3	52.0	51.0	+5.5
STCDA [37]	49.0	**52.6**	52.0	55.6	45.5	**52.5**	51.2	+5.7
MD-DMD	**50.3**	51.0	**56.0**	54.7	**47.3**	52.4	**52.0**	**+6.5**
Supervised-target	62.8	62.8	71.7	71.7	74.0	74.0	69.5	-

Table 2: Performance comparison on UCF-HMDB dataset. MM refers to Multi-Modal, U and H represent UCF and HMDB respectively.

Method	MM	U→H	H→U	Avg	Gain
Source-only [37] (R)		80.8	88.4	84.6	-
SAVA [8] (R)		82.2	91.2	86.7	+2.1
STCDA [37] (R)		81.9	91.9	86.9	+2.3
Source-only (R)		77.2	86.5	81.9	-
MixDANN (R)		77.5	86.5	82.0	+0.1
Supervised-target (R)		93.1	97.0	95.1	-
Source-only [37]	✓	82.8	89.8	86.3	-
MM-SADA [29]	✓	84.2	91.1	87.7	+1.4
CrossModal [23]	✓	84.7	92.8	88.8	+2.5
STCDA [37]	✓	83.1	92.1	87.6	+1.3
Source-only	✓	80.8	91.0	85.9	-
MD-DMD	✓	82.2	92.8	87.5	+1.6
Supervised-target	✓	98.8	95.0	96.9	-

achieves noticeable improvements from 48.6 to 52.0. Furthermore, MD-DMD manages to outperform the SOTAs MM-SADA [29], CrossModal [23] and STCDA [37] by 3.4%, 2.0% and 1.6% on average, respectively. Note that MD-DMD is simple and lightweight by adding two plug-and-play units, D and C_{mix}. Both units are shallow enough. In contrast, CrossModal [23] and STCDA [37] are both Contrastive-Learning (CL) based methods. They both contain two CL sub-tasks with properly designed sampling strategies to generate pseudo-label for unlabelled target data.

On UCF-HMDB datasets, we conduct experiments in two directions, UCF to HMDB (U→H) and HMDB to UCF (H→U). Table 2 is divided into two categories, *i.e.*, using single-modality RGB-only (RGB) and multi-modal (MM) for domain adaptation. Note that since our implementation does not meet the performance reported in [8, 23, 29, 37] on Source-only, we list our results and compare the improvements to Source-only separately. In a single modality, our proposed MixDANN has no obvious improvements. However, in the lower group, our complete MD-DMD framework for multi-modal domain adaptation gains a 1.6% increase over Source-only, which is higher than the 1.4% and 1.3% of MM-SADA [29] and STCDA [37], respectively.

4.4 Ablation Study

Our ablation studies investigate the individual impact of each component in our MD-DMD framework and the different configurations of the hyperparameters.

MixDANN. Firstly, we test the effectiveness of our MixDANN and the Mix-Samples Classification sub-task in every single modality on the EPIC-Kitchens dataset. The RGB and Flow modalities are trained individually without cross-modal alignment, illustrated in Table 3, where the C_{mix} indicates the mix-sample classifier. Specifically, for each modality, the first row reports the direct transfer results trained on labelled source data without any adaptation methods, and the second row is the results trivially using DANN as a baseline. The third row shows the effectiveness of our proposed MixDANN, which extends by DANN, which improves the performance by 8.9% in RGB and 1.8% in Flow compared with DANN. In MixDANN, the domain discriminator is forced not only to distinguish the inputs from source or target domains but also to predict the mix ratio of mix-samples. This process makes the feature extractor need to extract more domain-invariant features to confuse the discriminator. The last row shows the results of MixDANN with the mix-sample classification sub-task. There is a slight (1.1%) drop in RGB modality, but in Flow modality, there is a 2.4% improvement. Overall it brings improvement. Note that the C_{mix} is vital for conducting the direction of our proposed Dynamic-Modal-Distillation (DMD), which we reported in the next ablation. The difference between RGB and Flow demonstrates that Flow modality is more domain-invariant than RGB. The performance of Flow is 7.8% larger than RGB when directly transferring the model trained on source data to the target domain. Our MixDANN reduces this gap to 1.1% by 14.1% improvement in RGB and 7% in Flow. This phenomenon also demonstrates that the degrees of domain shift of modalities and the adaptability of sub-models are diverse, which motivates us to propose DMD.

MD-DMD. We further investigate the effect of different temperatures on Modal-Distillation in Eq 11, listed in Table 4. The results

Table 3: Ablation study of MixDANN and Mix-Prediction sub-task in each single modality.

Modality	Setting			D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	Avg	Gain
	DANN	MixDANN	C_{mix}								
RGB	✗	✗	✗	37.0	34.7	39.7	44.9	38.3	42.7	39.6	-
	✓	✗	✗	37.8	41.1	45.7	45.1	38.1	41.2	41.5	+1.9
	✓	✓	✗	**44.6**	**43.0**	**50.3**	46.4	42.0	**44.9**	**45.2**	**+5.6**
	✓	✓	✓	43.4	**43.0**	48.0	**46.9**	**43.0**	43.9	44.7	+5.1
Flow	✗	✗	✗	38.9	44.4	44.7	45.6	37.4	45.3	42.7	-
	✓	✗	✗	42.8	42.8	47.3	46.9	41.5	48.1	44.9	+2.2
	✓	✓	✗	42.5	46.0	49.7	48.3	40.5	**47.4**	45.7	+3.0
	✓	✓	✓	**45.3**	**46.9**	**50.5**	**49.7**	**40.8**	47.3	**46.8**	**+4.1**

Table 4: Performance of MD-DMD with different Temperature T.

Method	T	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	Avg	Gain
Source-only	-	42.5	44.3	42.0	56.3	41.2	46.5	45.5	-
MD-DMD	2	50.3	49.7	54.7	54.0	47.8	**52.4**	51.5	+6.0
	5	50.3	**51.0**	56.0	54.7	47.3	**52.4**	**52.0**	**+6.5**
	10	**50.6**	50.8	56.0	54.9	47.6	51.6	51.9	+6.4
	20	49.9	49.9	**56.9**	**55.1**	**48.4**	51.1	51.7	+6.2

show that the optimum distillation temperature is different. This may be due to the degrees of domain shift between domains being diverse. On average, the best result occurred when the temperature was 5. Note that even the worst average value of 51.5 ($T = 2$) also achieves SOTA results.

Table 5: Ablation study on different components' contributions to Teacher Score ($\mathcal{TS}$).

Setting			RGB	:	Flow	Accuracy
$\mathcal{L}_{D_{vanilla}}$	$\mathcal{L}_{D_{mix}}$	$\mathcal{L}_{MixCls}$				
✓	✗	✗	7619	:	8352	49.1
✓	✓	✗	7903	:	8075	49.2
✓	✓	✓	5290	:	10711	**50.6**

Teacher Score. We also examine the contributions of different components (especially the mix-sample classifier C_{mix}) in the Teacher Score $\mathcal{TS}$, which controls the direction of knowledge transfer in our proposed Modal-Distillation, listed in Table 5. The experiment is conducted on EPIC-Kitchens D2→D1, and the distillation temperature is set to 10, the $\mathcal{L}_{D_{vanilla}}$, $\mathcal{L}_{D_{mix}}$ and $\mathcal{L}_{MixCls}$ corresponding with the components DANN, MixDANN and C_{mix}, respectively. We report the ratios of the number of RGB playing the role of Teacher Modality with the number of Flow as Teacher over the 16000 training steps. Results show that the C_{mix} dominates the direction of Modal-Distillation, which can make $\mathcal{TS}$ reflecting the adaptability of single modality sub-models more accurate. According to our observation, numerically, $\mathcal{L}_{MixCls}$ varies in a larger range and has a greater impact on the teacher score. And the RGB modality converges faster than Flow in this mix-sample classification task, causing the loss value of Flow to be almost always greater than RGB before convergence. This reflects the inherent domain-invariant of the Flow modality, which makes it more difficult to distinguish than RGB. In the adversarial training tasks, the differences in loss values are not so obvious. Note that since teacher scores may sometimes be equal, the sum of the teaching times may not necessarily be equal to the total training steps of 16000.

We also have observed the changes of teacher modality during training. In the beginning, the Optical Flow almost always plays the role of teacher to guide RGB modality. Then, RGB begins to teach Flow gradually. When it comes to convergence, the frequencies of RGB and Flow as teachers are similar.

5 CONCLUSION

We have presented MixDANN and Dynamic-Modal-Distillation (MD-DMD) framework for video domain adaptation, which makes full use of the temporal dimension of video inputs for adversarial learning in a single modality and explored the adaptability difference between modalities for multi-modal domain adaptation. We produce mix-samples to enhance adversarial strength, dynamically measure the adaptability of each modality and let modalities guide and reinforce each other alternatively for better multi-modal performance. Experimental superior results over SOTAs [8, 23, 29, 37] demonstrate the effectiveness of our framework.

ACKNOWLEDGMENTS

This work was supported in part by Shanghai Science and Technology Program (No. 21JC1400600), and Exploratory Research Project (2022PG0AN01). Y.-G. Jiang was sponsored in part by "Shuguang

Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (No. 20SG01).

REFERENCES

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.

[2] Taichi Asami, Ryo Masumura, Yoshikazu Yamaguchi, Hirokazu Masataki, and Yushi Aono. 2017. Domain adaptation of dnn acoustic models using knowledge distillation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5185–5189.

[3] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. 2019. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11457–11466.

[4] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2229–2238.

[5] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[6] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. 2019. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6321–6330.

[7] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. 2019. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1791–1800.

[8] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. 2020. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*. Springer, 678–695.

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 720–736.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2020. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2020), 4125–4141.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[13] Muhammad Ghifary, W Bastiaan Kleijn, and Mengjie Zhang. 2014. Domain adaptive neural networks for object recognition. In *Pacific Rim international conference on artificial intelligence*. Springer, 898–904.

[14] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*. Springer, 597–613.

[15] Eric Granger, Madhu Kiran, Jose Dolz, Louis-Antoine Blais-Morin, et al. 2020. Joint progressive knowledge distillation and unsupervised domain adaptation. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[16] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. 2022. Group Contextualization for Video Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 928–938.

[17] Zhenwei He and Lei Zhang. 2019. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6668–6677.

[18] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[19] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.

[20] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. 2018. Deep Domain Adaptation in Action Space.. In *BMVC*, Vol. 2. 5.

[21] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2012), 221–231.

[22] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[23] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. 2021. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13618–13627.

[24] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*. IEEE, 2556–2563.

[25] Jinyu Li, Michael L Seltzer, Xi Wang, Rui Zhao, and Yifan Gong. 2017. Large-scale domain adaptation via teacher-student learning. *arXiv preprint arXiv:1708.05466* (2017).

[26] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. 2018. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition* 80 (2018), 109–117.

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3202–3211.

[28] Jianming Lv, Kaijie Liu, and Shengfeng He. 2021. Differentiated Learning for Multi-Modal Domain Adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1322–1330.

[29] Jonathan Munro and Dima Damen. 2020. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 122–132.

[30] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. 2021. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1094–1103.

[31] Mauricio Orbes-Arteainst, Jorge Cardoso, Lauge Sørensen, Christian Igel, Sebastien Ourselin, Marc Modat, Mads Nielsen, and Akshay Pai. 2019. Knowledge distillation for semi-supervised domain adaptation. In *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*. Springer, 68–76.

[32] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. 2020. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11815–11822.

[33] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.

[34] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. 2021. Contrast and Mix: Temporal Contrastive Video Domain Adaptation with Background Mixing. *Advances in Neural Information Processing Systems* 34 (2021).

[35] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3723–3732.

[36] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems* 27 (2014).

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

[38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[39] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*. Springer, 443–450.

[40] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. 2019. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825* (2019).

[41] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[42] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.

[43] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7167–7176.

[44] Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312 (2018), 135–153.

[45] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. 2021. Bevt: Bert pretraining of video transformers. *arXiv preprint arXiv:2112.01529* (2021).

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.

[47] Zuxuan Wu, Hengduo Li, Caiming Xiong, Yu-Gang Jiang, and Larry Steven Davis. 2020. A dynamic frame selection framework for fast video recognition. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence (2020).

[48] Zuxuan Wu, Hengduo Li, Yingbin Zheng, Caiming Xiong, Yu-Gang Jiang, and Larry S. Davis. 2021. A Coarse-to-Fine Framework for Resource Efficient Video Recognition. *IJCV* 129, 11 (2021), 2965–2977.

[49] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[50] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. 2021. Token shift transformer for video classification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 917–925.

[51] Jing Zhang, Wanqing Li, and Philip Ogunbona. 2017. Joint geometrical and statistical alignment for visual domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1859–1867.

[52] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. 2015. Supervised representation learning: Transfer learning with deep autoencoders. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.