This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Temporal Attentive Alignment for Large-Scale Video Domain Adaptation

Min-Hung Chen[1]* Zsolt Kira[1] Ghassan AlRegib[1] Jaekwon Yoo[2] Ruxin Chen[2] Jian Zheng[3]*
[1]Georgia Institute of Technology [2]Sony Interactive Entertainment LLC [3]Binghamton University

Abstract

Although various image-based domain adaptation (DA) techniques have been proposed in recent years, domain shift in videos is still not well-explored. Most previous works only evaluate performance on small-scale datasets which are saturated. Therefore, we first propose two large-scale video DA datasets with much larger domain discrepancy: ***UCF-HMDB$_{full}$*** *and* ***Kinetics-Gameplay****. Second, we investigate different DA integration methods for videos, and show that simultaneously aligning and learning temporal dynamics achieves effective alignment even without sophisticated DA methods. Finally, we propose* ***Temporal Attentive Adversarial Adaptation Network (TA^{3}N)****, which explicitly attends to the temporal dynamics using domain discrepancy for more effective domain alignment, achieving state-of-the-art performance on four video DA datasets (e.g. 7.9% accuracy gain over "Source only" from 73.9% to 81.8% on "HMDB → UCF", and 10.3% gain on "Kinetics → Gameplay"). The code and data are released at* http://github.com/cmhungsteve/TA3N.

1. Introduction

Domain adaptation (DA) [29] has been studied extensively in recent years [4] to address the *domain shift* problem [32, 30], which means the models trained on source labeled dataset do not generalize well to target datasets and tasks. DA is categorized in terms of the availability of annotations in the target domain. In this paper, we focus on the harder unsupervised DA problem, which requires training models that can generalize to target samples without access to any target labels. While many unsupervised DA approaches are able to diminish the distribution gap between source and target domains while learning discriminative deep features [22, 24, 9, 10, 21, 20, 34], most methods have been developed only for images and not videos.

Furthermore, unlike image-based DA work, there do not exist well-organized datasets to evaluate and benchmark the performance of DA algorithms for videos. The most common datasets are *UCF-Olympic* and *UCF-HMDB$_{small}$* [39, 46, 15], which have only a few overlapping categories between source and target domains. This introduces limited domain discrepancy so that a deep CNN architecture can achieve nearly perfect performance even without any DA method (details in Section 5.2 and Table 2). Therefore, we propose two larger-scale datasets to investigate video DA: 1) *UCF-HMDB$_{full}$*: We collect 12 overlapping categories between UCF101 [38] and HMDB51 [18], which is around three times larger than both UCF-Olympic and UCF-HMDB$_{small}$, and contains larger domain discrepancy (details in Section 5.2 and Tables 3 and 4). 2) *Kinetics-Gameplay*: We collect from several currently

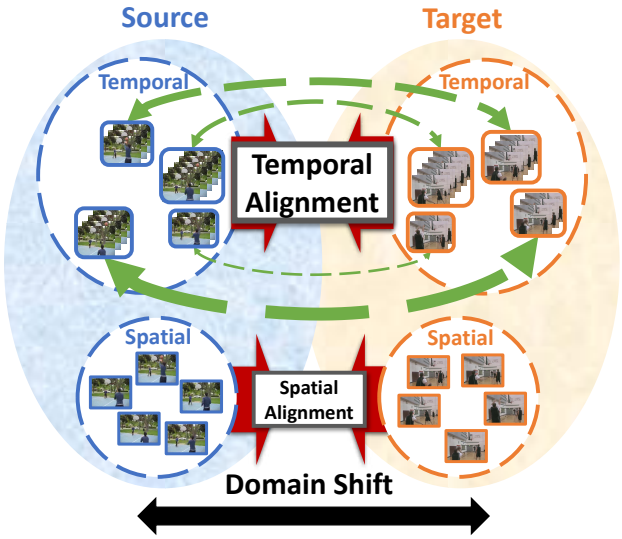


Figure 1: An overview of proposed TA^{3}N for video DA. In addition to spatial discrepancy between frame images, videos also suffer from temporal discrepancy between sets of time-ordered frames that contain multiple local temporal dynamics with different contributions to the overall domain shift, as indicated by the thickness of green dashed arrows. Therefore, we propose to focus on aligning the temporal dynamics which have higher domain discrepancy using a learned attention mechanism to effectively align the temporal-embedded feature space for videos. Here we use the action *basketball* as the example.

*Work partially done as a SIE intern

popular video games with 30 overlapping categories with Kinetics-600 [17, 2]. This dataset is much more challenging than UCF-HMDB$_{full}$ due to the significant domain shift between the distributions of virtual and real data.

Videos can suffer from domain discrepancy along both the spatial and temporal directions, bringing the need of alignment for embedded feature spaces along both directions, as shown in Figure 1. However, most DA approaches have not explicitly addressed the domain shift problem in the temporal direction. Therefore, we first investigate different DA integration methods for video classification and show that: 1) aligning the features that encode temporal dynamics outperforms aligning only spatial features. 2) to effectively align domains spatio-temporally, *which features* to align is more important than *what DA approaches* to use. To support our claims, we then propose *Temporal Adversarial Adaptation Network (TA^{2}N)*, which simultaneously aligns and learns temporal dynamics, outperforming other approaches which naively apply more sophisticated image-based DA methods for videos.

The temporal dynamics in videos can be represented as a combination of multiple local temporal features corresponding to different motion characteristics. Not all of the local temporal features equally contribute to the overall domain shift. We want to focus more on aligning those which have high contribution to the overall domain shift, such as the local temporal features connected by thicker green arrows shown in Figure 1. Therefore, we propose **Temporal Attentive Adversarial Adaptation Network (TA^{3}N)** to explicitly attend to the temporal dynamics by taking into account the domain distribution discrepancy. In this way, the temporal dynamics which contribute more to the overall domain shift will be focused on, leading to more effective temporal alignment. TA^{3}N achieves state-of-the-art performance on all four investigated video DA datasets.

In summary, our contributions are three-fold:

1. **Video DA Dataset Collection**: We collect two large-scale video DA datasets, *UCF-HMDB$_{full}$* and *Kinetics-Gameplay*, to investigate the domain discrepancy problem across videos, which is an under-explored research problem. To our knowledge, they are by far the largest datasets for video DA problems.

2. **Feature Alignment Exploration for Video DA**: We investigate different DA integration approaches for videos and provide a strategy to effectively align domains spatio-temporally for videos by aligning temporal relation features. We propose this simple but effective approach, *TA^{2}N*, to demonstrate the importance of determining *what* to align over the DA method to use.

3. **Temporal Attentive Adversarial Adaptation Network (TA^{3}N)**: We propose *TA^{3}N*, which simultaneously aligns domains, encodes temporal dynamics into video representations, and attends to representations with domain distribution discrepancy. TA^{3}N achieves state-of-the-art performance on both small- and large-scale cross-domain video datasets.

2. Related Works

Video Classification. With the rise of deep convolutional neural networks (CNNs), recent work for video classification mainly aims to learn compact spatio-temporal representations by leveraging CNNs for spatial information and designing various architectures to exploit temporal dynamics [16]. In addition to separating spatial and temporal learning, some works propose different architectures to encode spatio-temporal representations with consideration of the trade-off between performance and computational cost [41, 3, 31, 42]. Another branch of work utilizes optical flow to compensate for the lack of temporal information in raw RGB frames [37, 7, 44, 3, 26]. Moreover, some works extract temporal dependencies between frames for video tasks by utilizing recurrent neural networks (RNNs) [5], attention [25, 27] and relation modules [51]. Note that we focus on attending to the temporal dynamics to effectively align domains and we consider other modalities, e.g. optical flow, to be complementary to our method.

Domain Adaptation. Most recent DA approaches are based on deep learning architectures designed for addressing the domain shift problems given the fact that the deep CNN features without any DA method outperform traditional DA methods using hand-crafted features [6]. Most DA approaches follow the two-branch (source and target) architecture, and aim to find a common feature space between the source and target domains. The models are therefore optimized with a combination of *classification* and *domain* losses [4].

One of the main classes of methods used is *Discrepancy-based DA*, whose metrics are designed to measure the distance between source and target feature distributions, including variations of maximum mean discrepancy (MMD) [22, 23, 48, 47, 24] and the CORAL function [40]. By diminishing the distance of distributions, discrepancy-based DA methods reduce the gap across domains. Another common method, *Adversarial-based DA*, adopts a similar concept as GANs [11] by integrating domain discriminators into the architectures. Through the adversarial objectives, the discriminators are optimized to classify different domains, while the feature extractors are optimized in the opposite direction. ADDA [43] uses an inverted label GAN loss to split the optimization into two parts: one for the discriminator and the other for the generator. In contrast, the gradient reversal layer (GRL) is used in some work [9, 10, 49] to invert the gradients so that the discriminator and generator are optimized simultaneously. Additionally, *Normalization-based DA* [21, 20] adapts batch nor-

malization [14] to DA problems by calculating two separate statistics, representing source and target, for normalization. Furthermore, *Ensemble-based DA* [8, 33, 34, 19] builds a target branch ensemble by incorporating multiple target branches. Recently, TADA [45] adopts the attention mechanism to adapt the transferable regions. We extend these concepts to spatio-temporal domains, aiming to attend to the important parts of temporal dynamics for alignment.

Video Domain Adaptation. Unlike image-based DA, video-based DA is still an under-explored area. Only a few works focus on small-scale video DA with only a few overlapping categories [39, 46, 15]. [39] improves the domain generalizability by decreasing the effect of the background. [46] maps source and target features to a common feature space using shallow neural networks. AMLS [15] adapts pre-extracted C3D [41] features on a Grassmann manifold obtained using PCA. However, the datasets used in the above works are too small to have enough domain shift to evaluate DA performance. Therefore, we propose two larger cross-domain datasets *UCF-HMDB*$_{full}$ and *Kinetics-Gameplay*, and provide benchmarks with different baseline approaches. Recently, TSRNet [50] transfers knowledge for action localization using MMD, but only aligns the video-level features. Instead, our *TA*^{3}*N* simultaneously attends, aligns, and encodes temporal dynamics into video features.

3. Technical Approach

We first introduce our baseline model which simply extends image-base DA for videos using the temporal pooling mechanism (Section 3.1). And then we investigate better ways to incorporate temporal dynamics for video DA (Section 3.2), and describe our final proposed method with the domain attention mechanism (Section 3.3).

3.1. Baseline Model

Given the recent success of large-scale video classification using CNNs [16], we build our baseline on such architectures, as shown in the lower part of Figure 2.

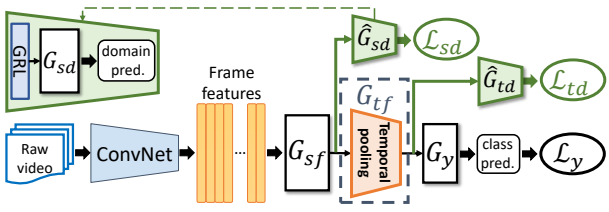


Figure 2: Baseline architecture (TemPooling) with the adversarial discriminators $\hat{G}_{sd}$ and $\hat{G}_{td}$. $\mathcal{L}_y$ is the class prediction loss, and $\mathcal{L}_{sd}$ and $\mathcal{L}_{td}$ are the domain losses. See the detailed architecture in the supplementary material.

We first feed the input video $X_i = \{x_i^1, x_i^2, ..., x_i^K\}$ extracted from ResNet [12] pre-trained on ImageNet into our model, where x_i^j is the jth frame-level feature representation of the ith video. The model can be divided into two parts: 1) *Spatial module* $G_{sf}(.;\theta_{sf})$, which consists of multilayer perceptrons (MLP) that aims to convert the general-purpose feature vectors into task-driven feature vectors, where the task is video classification in this paper; 2) *Temporal module* $G_{tf}(.;\theta_{tf})$ aggregates the frame-level feature vectors to form a single video-level feature vector for each video. In our baseline architecture, we conduct mean-pooling along the temporal direction to generate video-level feature vectors, and note it as *TemPooling*. Finally, another fully-connected layer $G_y(.;\theta_y)$ converts the video-level features into the final predictions, which are used to calculate the class prediction loss $\mathcal{L}_y$.

Similar to image-based DA problems, the baseline approach is not able to generalize to data from different domains due to domain shift. Therefore, we integrate TemPooling with the unsupervised DA method inspired by one of the most popular adversarial-based approaches, DANN [9, 10]. The main idea is to add additional domain classifiers $G_d(.;\theta_d)$, to discriminate whether the data is from the source or target domain. Before back-propagating the gradients to the main model, a gradient reversal layer (GRL) is inserted between G_d and the main model to invert the gradient, as shown in Figure 2. During adversarial training, the parameters θ_{sf} are learned by maximizing the domain discrimination loss $\mathcal{L}_d$, and parameters θ_d are learned by minimizing $\mathcal{L}_d$ with the domain label d. Therefore, the feature generator G_f will be optimized to gradually align the feature distributions between the two domains.

In this paper, we note the *Adversarial Discriminator* $\hat{G}_d$ as the combination of a gradient reversal layer (GRL) and a domain classifier, and insert $\hat{G}_d$ into TemPooling in two ways: 1) $\hat{G}_{sd}$: show how directly applying image-based DA approaches can benefit video DA; 2) $\hat{G}_{td}$: indicate how DA on temporal-dynamics-encoded features benefits video DA.

The prediction loss $\mathcal{L}_y$, spatial domain loss $\mathcal{L}_{sd}$ and temporal domain loss $\mathcal{L}_{td}$ can be expressed as follows (ignoring all the parameter symbols through the paper to save space):

$$\mathcal{L}_y^i = L_y(G_y(G_{tf}(G_{sf}(X_i))), y_i) \quad (1)$$

$$\mathcal{L}_{sd}^i = \frac{1}{K}\sum_{j=1}^{K} L_d(G_{sd}(G_{sf}(x_i^j)), d_i) \quad (2)$$

$$\mathcal{L}_{td}^i = L_d(G_{td}(G_{tf}(G_{sf}(X_i))), d_i) \quad (3)$$

where K is the number of frames sampled from each video. L is the cross entropy loss function.

The overall loss can be expressed as follows:

$$\mathcal{L} = \frac{1}{N_S}\sum_{i=1}^{N_S}\mathcal{L}_y^i - \frac{1}{N_{S\cup T}}\sum_{i=1}^{N_{S\cup T}}(\lambda_s\mathcal{L}_{sd}^i + \lambda_t\mathcal{L}_{td}^i) \quad (4)$$

where N_S equals the number of source data, and $N_{S\cup T}$ equals the number of all data. λ_s and λ_t is the trade-off weighting for spatial and temporal domain loss.

3.2. Integration of Temporal Dynamics with DA

One main drawback of directly integrating image-based DA approaches into our baseline architecture is that the feature representations learned in the model are mainly from the spatial features. Although we implicitly encode the temporal information by the temporal pooling mechanism, the relation between frames is still missing. Therefore, we would like to address two questions: 1) *Does the video DA problem benefit from encoding temporal dynamics into features?* 2) *Instead of only modifying feature encoding methods, how can DA be further integrated while encoding temporal dynamics into features?*

To answer the first question, given the fact that humans can recognize actions by reasoning the observations across time, we propose the *TemRelation* architecture by replacing the temporal pooling mechanism with the Temporal Relation module, which is modified from [36, 51], as shown in Figure 4.

The n-frame temporal relation is defined by the function:

$$R_n(V_i) = \sum_m g_{\phi^{(n)}}((V_i^n)_m) \tag{5}$$

where $(V_i^n)_m = \{v_i^a, v_i^b, ...\}_m$ is the mth set of frame-level representations from n temporal-ordered sampled frames. a and b are the frame indices. We fuse the feature vectors that are time-ordered with the function $g_{\phi^{(n)}}$, which is an MLP with parameters $\phi^{(n)}$. To capture temporal relations at multiple time scales, we sum up all the n-frame relation features into the final video representation. In this way, the temporal dynamics are explicitly encoded into features. We then insert $\hat{G}_d$ into TemRelation as we did for TemPooling.

Although aligning temporal-dynamic-encoded features benefits video DA, feature encoding and DA are still two separate processes, leading to sub-optimal DA performance. Therefore, we address the second question by proposing **Temporal Adversarial Adaptation Network (TA^{2}N)**, which explicitly integrates $\hat{G}_d$ *inside* the Temporal module to align the model across domains while learning temporal dynamics. Specifically, we integrate each n-frame relation with a corresponding relation discriminator $\hat{G}_{rd}^n$ because different n-frame relations represent different temporal characteristics, which correspond to different parts of actions. The relation domain loss $\mathcal{L}_{rd}$ can be expressed as follows:

$$\mathcal{L}_{rd}^i = \frac{1}{K-1}\sum_{n=2}^{K} L_d(G_{rd}^n(R_n(G_{sf}(X_i))), d_i) \tag{6}$$

The experimental results show that our integration strategy can effectively align domains spatio-temporally for videos, and outperform those which are extended from sophisticated DA approaches although TA^{2}N is adopted from a simpler DA method (DANN) (see details in Tables 3 to 5).

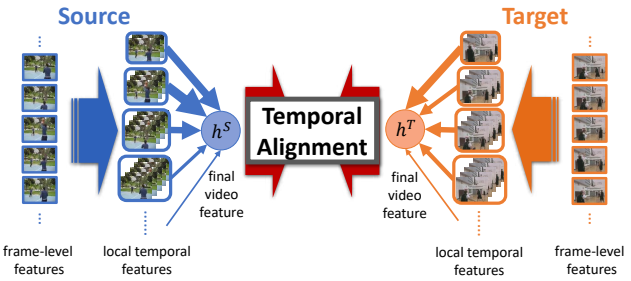


Figure 3: The domain attention mechanism in TA^{3}N. Thicker arrows correspond to larger attention weights.

3.3. Temporal Attentive Alignment for Videos

The final video representation of TA^{2}N is generated by aggregating multiple local temporal features. Although aligning temporal features across domains benefits video DA, not all the features are equally important to align. In order to effectively align overall temporal dynamics, we want to focus more on aligning the local temporal features which have larger domain discrepancy. Therefore, we represent the final video representation as a combination of local temporal features with different attention weighting, as shown in Figure 3, and aim to attend to features of interest that are domain discriminative so that the DA mechanism can focus on aligning those features. The main question becomes: *How to incorporate domain discrepancy for attention?*

To address this, we propose **Temporal Attentive Adversarial Adaptation Network (TA^{3}N)**, as shown in Figure 4, by introducing the *domain attention* mechanism, which utilize the entropy criterion to generate the domain attention value for each n-frame relation feature as below:

$$w_i^n = 1 - H(\hat{d}_i^n) \tag{7}$$

where $\hat{d}_i^n$ is the output of G_{rd}^n for the ith video. $H(p) = -\sum_k p_k \cdot \log(p_k)$ is the entropy function to measure uncertainty. w_i^n increases when $H(\hat{d}_i^n)$ decreases, which means the domains can be distinguished well. We also add a residual connection for more stable optimization. Therefore, the final video feature representation h_i generated from attended local temporal features, which are learned by local temporal modules $G_{tf}^{(n)}$, can be expressed as:

$$h_i = \sum_{n=2}^{K}(w_i^n + 1) \cdot G_{tf}^{(n)}(G_{sf}(X_i)) \tag{8}$$

Finally, we add the minimum entropy regularization to refine the classifier adaptation. However, we only want to minimize the entropy for the videos that are similar across domains. Therefore, we attend to the videos which have low domain discrepancy, so that we can focus more on minimizing the entropy for these videos. The attentive entropy

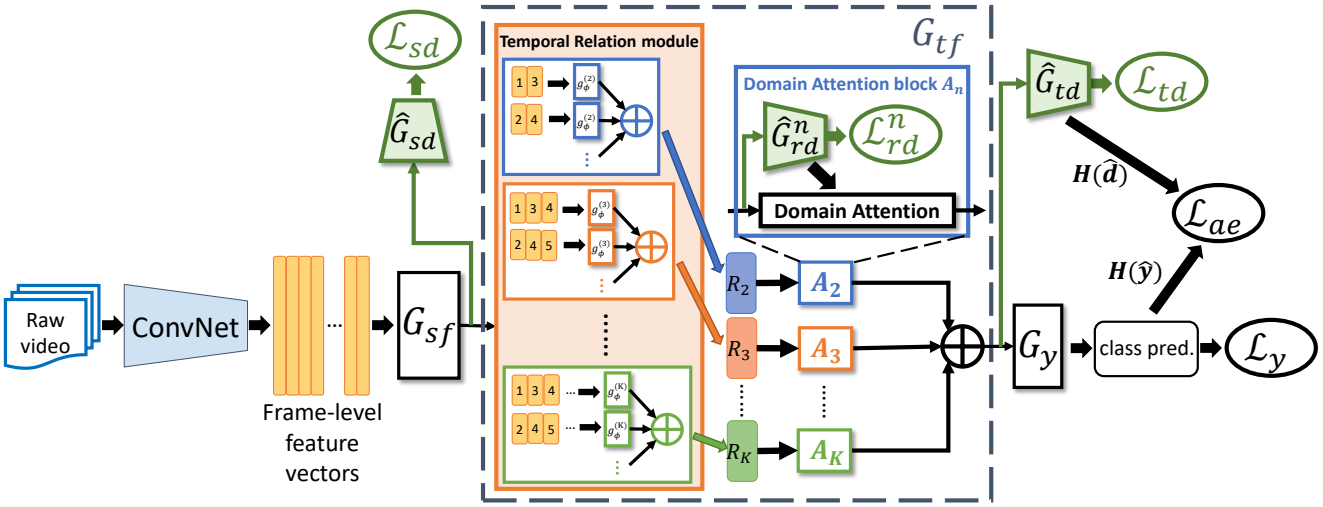


Figure 4: The overall architecture of the proposed Temporal Attentive Adversarial Adaptation Network (TA^{3}N). In the temporal relation module, time-ordered frames are used to generate K-1 relation feature representations $\boldsymbol{R} = \{R_2, ..., R_K\}$, where R_n corresponds to the n-frame relation (the numbers in this figure are examples of time indices). After attending with the domain predictions from relation discriminators G^n_{rd}, the relation features are summed up to the final video representation. The attentive entropy loss $\mathcal{L}_{ae}$, which is calculated by domain entropy $H(\hat{d})$ and class entropy $H(\hat{y})$, aims to enhance the certainty of those videos that are more similar across domains. See the detailed architecture in the supplementary material.

loss $\mathcal{L}_{ae}$ can be expressed as follows:

$$\mathcal{L}^i_{ae} = (1 + H(\hat{d}_i)) \cdot H(\hat{y}_i) \tag{9}$$

where $\hat{d}_i$ and $\hat{y}_i$ is the output of G_{td} and G_y, respectively. We also adopt the residual connection for stability.

By combining Equations (1) to (3), (6) and (9), and replacing G_{sf} and G_{tf} with h_i by Equation (8), the overall loss of TA^{3}N can be expressed as follows:

$$\begin{aligned}\mathcal{L} = &\frac{1}{N_S}\sum_{i=1}^{N_S}\mathcal{L}^i_y + \frac{1}{N_{S\cup T}}\sum_{i=1}^{N_{S\cup T}}\gamma\mathcal{L}^i_{ae} \\ &- \frac{1}{N_{S\cup T}}\sum_{i=1}^{N_{S\cup T}}(\lambda^s\mathcal{L}^i_{sd} + \lambda^r\mathcal{L}^i_{rd} + \lambda^t\mathcal{L}^i_{td})\end{aligned} \tag{10}$$

where λ^s, λ^r and λ^t is the trade-off weighting for each domain loss. γ is the weighting for the attentive entropy loss. All the weightings are chosen via grid search.

Our proposed TA^{3}N and TADA [45] both utilize entropy functions for attention but with different perspectives. TADA aims to focus on the foreground objects for image DA, while TA^{3}N aims to find important and discriminative parts of temporal dynamics to align for video DA.

4. Datasets

There are very few benchmark datasets for video DA, and only small-scale datasets have been widely used [39, 46, 15]. Therefore, we specifically create two cross-domain datasets to evaluate the proposed approaches for the video DA problem, as shown in Table 1. For more details about the datasets, please refer to the supplementary material.

4.1. UCF-HMDB$_{full}$

We extend UCF-HMDB$_{small}$ [39], which only selects 5 visually highly similar categories, by collecting all of the relevant and overlapping categories between UCF101 [38] and HMDB51 [18], which results in 12 categories. We follow the official split method to separate training and validation sets. This dataset, **UCF-HMDB$_{full}$**, includes more than 3000 video clips, which is around 3 times larger than UCF-HMDB$_{small}$ and UCF-Olympic.

4.2. Kinetics-Gameplay

In addition to real-world videos, we are also interested in virtual-world videos for DA. While there are more than ten real-world video datasets, there is a limited number of virtual-world datasets for video classification. It is mainly because rendering realistic human actions using game engines requires gaming graphics expertise which is time-consuming. Therefore, we create the *Gameplay* dataset by collecting gameplay videos from currently popular video games, *Detroit: Become Human* and *Fortnite*, to build our own video dataset for the virtual domain. For the real domain, we use one of the largest public video datasets *Kinetics-600* [17, 2]. We follow the closed-set DA setting [30] to select 30 overlapping categories between the

	UCF-HMDB$_{small}$	UCF-Olympic	UCF-HMDB$_{full}$	Kinetics-Gameplay
length (sec.)	1 - 21	1 - 39	1 - 33	1 - 10
class #	5	6	12	30
video #	1171	1145	3209	49998

Table 1: The comparison of the cross-domain video datasets.

Kinetics-600 and Gameplay datasets to build the **Kinetics-Gameplay** dataset with both domains, including around 50K video clips. See the supplementary material for the complete statistics and example snapshots.

5. Experiments

We therefore evaluate DA approaches on four datasets: UCF-Olympic, UCF-HMDB$_{small}$, UCF-HMDB$_{full}$ and Kinetics-Gameplay.

5.1. Experimental Setup

UCF-Olympic and **UCF-HMDB$_{small}$**. First, we evaluate our approaches on UCF-Olympic and UCF-HMDB$_{small}$, and compare with all other works that also evaluate on these two datasets [39, 46, 15]. We follow the default settings, but the method to split the UCF video clips into training and validations sets is not specified in these papers, so we follow the official split method from UCF101 [38].

UCF-HMDB$_{full}$ and **Kinetics-Gameplay**. For the self-collected datasets, we follow the common experimental protocol of unsupervised DA [30]: the training data consists of labeled data from the source domain and unlabeled data from the target domain, and the validation data is all from the target domain. However, unlike most of the image DA settings, our training and validation data in both domains are separate to avoid potentially overfitting while aligning different domains. To compare with image-based DA approaches, we extend several state-of-the-art methods [10, 24, 20, 34] for video DA with our TemPooling and TemRelation architectures, as shown in Tables 3 to 5. The difference between the "Target only" and "Source only" settings is the domain used for training. The "Target only" setting can be regarded as the upper bound without domain shift while the "Source only" setting shows the lower bound which directly applies the model trained with source data to the target domain without modification. See supplementary materials for full implementation details.

5.2. Experimental Results

UCF-Olympic and **UCF-HMDB$_{small}$**. In these two datasets, our approach outperforms all the previous methods by at least 6.5% absolute difference (98.15% - 91.60%) on the "U → O" setting, and 9% difference (99.33% - 90.25%) on the "U → H" setting, as shown in Table 2.

Source → Target	U → O	O → U	U → H	H → U
W. Sultani et al. [39]	33.33	47.91	68.70	68.67
T. Xu et al. [46]	87.00	75.00	82.00	82.00
AMLS (GFK) [15]†	84.65	86.44	89.53	95.36
AMLS (SA) [15]†	83.92	86.07	90.25	94.40
DAAA [15]†‡	91.60	89.96	-	-
TemPooling	96.30	87.08	98.67	97.35
TemPooling + DANN [10]	**98.15**	90.00	**99.33**	98.41
Ours (TA^{2}N)	**98.15**	91.67	**99.33**	**99.47**
Ours (TA^{3}N)	**98.15**	**92.92**	**99.33**	**99.47**

Table 2: The accuracy (%) for the state-of-the-art work on UCF-Olympic and UCF-HMDB$_{small}$ (U: UCF, O: Olympic, H: HMDB). †We only show their results which are fine-tuned with source data for fair comparison. Please refer to the supplementary material for more details. ‡[15] did not test DAAA on UCF-HMDB$_{small}$.

These results also show that the performance on these datasets is saturated. With a strong CNN as the backbone architecture, even our baseline architecture TemPooling can achieve high accuracy without any DA method (e.g. 96.3% for "U → O"). This suggests that these two datasets are not enough to evaluate more sophisticated DA approaches, so larger-scale datasets for video DA are needed.

UCF-HMDB$_{full}$. We then evaluate our approaches and compare with other image-based DA approaches on the UCF-HMDB$_{full}$ dataset, as shown in Tables 3 and 4. The accuracy difference between "Target only" and "Source only" indicates the *domain gap*. The gaps for the HMDB dataset are 11.11% for TemRelation and 10.28% for TemPooling (see Table 3), and the gaps for the UCF dataset are 21.01% for TemRelation and 17.16% for TemPooling (see Table 4). It is worth noting that the "Source only" accuracy of our baseline architecture (TemPooling) on UCF-HMDB$_{full}$ is much lower than UCF-HMDB$_{small}$ (e.g. 28.39 lower for "U → H"), which implies that UCF-HMDB$_{full}$ contains much larger domain discrepancy than UCF-HMDB$_{small}$. The value "Gain" is the difference from the "Source only" accuracy, which directly indicates the effectiveness of the DA approaches. We now answer the two questions for video DA in Section 3.2 (see Tables 3 and 4):

1. *Does the video DA problem benefit from encoding temporal dynamics into features?*

   From Tables 3 and 4, we see that for the same DA method, TemRelation outperforms TemPooling in

Temporal Module	TemPooling		TemRelation	
	Acc.	Gain	Acc.	Gain
Target only	80.56	-	82.78	-
Source only	70.28	-	71.67	-
DANN [10]	71.11	0.83	75.28	3.61
JAN [24]	71.39	1.11	74.72	3.05
AdaBN [20]	75.56	5.28	72.22	0.55
MCD [34]	71.67	1.39	73.89	2.22
Ours (TA^{2}N)	N/A	-	77.22	5.55
Ours (TA^{3}N)	N/A	-	**78.33**	**6.66**

Table 3: The comparison of accuracy (%) with other approaches on UCF-HMDB$_{full}$ (U $\rightarrow$ H). Gain represents the absolute difference from the "Source only" accuracy. TA^{2}N and TA^{3}N are based on the TemRelation architecture, so they are not applicable to TemPooling.

Temporal Module	TemPooling		TemRelation	
	Acc.	Gain	Acc.	Gain
Target only	92.12	-	94.92	-
Source only	74.96	-	73.91	-
DANN [10]	75.13	0.17	76.36	2.45
JAN [24]	80.04	5.08	79.69	5.79
AdaBN [20]	76.36	1.40	77.41	3.51
MCD [34]	76.18	1.23	79.34	5.44
Ours (TA^{2}N)	N/A	-	80.56	6.66
Ours (TA^{3}N)	N/A	-	**81.79**	**7.88**

Table 4: The comparison of accuracy (%) with other approaches on UCF-HMDB$_{full}$ (H $\rightarrow$ U).

most cases, especially for the gain value. For example, "TemPooling+DANN" reaches 0.83% absolute accuracy gain on the "U $\rightarrow$ H" setting and 0.17% gain on the "H $\rightarrow$ U" setting while "TemRelation+DANN" reaches 3.61% gain on "U $\rightarrow$ H" and 2.45% gain on "H $\rightarrow$ U". This means that applying DA approaches to the video representations which encode the temporal dynamics improves the overall performance for cross-domain video classification.

2. *How to further integrate DA while encoding temporal dynamics into features?*

   Although integrating TemRelation with image-based DA approaches generally has better alignment performance than the baseline (TemPooling), feature encoding and DA are still two separate processes. The alignment happens only before and after the temporal dynamics are encoded in features. In order to explicitly force alignment of the temporal dynamics across domains, we propose TA^{2}N, which reaches 77.22% (5.55% gain) on "U $\rightarrow$ H" and 80.56% (6.66% gain) on "H $\rightarrow$ U". Tables 3 and 4 show that although TA^{2}N is adopted from a simple DA method (DANN), it still outperforms other approaches which are extended from more sophisticated DA methods but do not follow our strategy.

Finally, with the domain attention mechanism, our proposed **TA^{3}N** reaches 78.33% (6.66% gain) on "U $\rightarrow$ H" and 81.79% (7.88% gain) on "H $\rightarrow$ U", achieving state-of-the-art performance on UCF-HMDB$_{full}$ in terms of accuracy and gain, as shown in Tables 3 and 4.

Kinetics-Gameplay. Kinetics-Gameplay is much more challenging than UCF-HMDB$_{full}$ because the data is from real and virtual domains, which have more severe domain shifts. Here we only utilize TemRelation as our backbone architecture since it is proved to outperform TemPooling on UCF-HMDB$_{full}$. Table 5 shows that the accuracy gap between "Source only" and "Target only" is 47.27%, which is more than twice the number in UCF-HMDB$_{full}$. In this dataset, TA^{3}N also outperforms all the other DA approaches by increasing the "Source only " accuracy from 17.22% to 27.50%.

	Acc.	Gain
Target only	64.49	-
Source only	17.22	-
DANN [10]	20.56	3.34
JAN [24]	18.16	0.94
AdaBN [20]	20.29	3.07
MCD [34]	19.76	2.54
Ours (TA^{2}N)	24.30	7.08
Ours (TA^{3}N)	**27.50**	**10.28**

Table 5: The comparison of accuracy (%) with other approaches on Kinetics-Gameplay.

5.3. Ablation Study and Analysis

Integration of $\hat{G}_d$. We use UCF-HMDB$_{full}$ to investigate the performance for integrating $\hat{G}_d$ in different positions. There are three ways to insert the adversarial discriminator into our architectures, where each corresponds to different feature representations, leading to three types of discriminators $\hat{G}_{sd}$, $\hat{G}_{td}$ and $\hat{G}_{rd}$, which are shown in Figure 4 and the full experimental results are shown in Table 6. For the TemRelation architecture, the accuracy of utilizing $\hat{G}_{td}$ shows better performance than utilizing $\hat{G}_{sd}$ (averagely 0.58% absolute gain improvement across two tasks), while the accuracies are the same for TemPooling. This means that the temporal relation module can encode temporal dynamics that help the video DA problem, but temporal pooling cannot. Utilizing the relation discriminator $\hat{G}_{rd}$ can further improve the performance (0.92% improvement) since we simultaneously align and learn the temporal dynamics across domains. Finally, by combining all three discriminators, TA^{2}N improves even more (4.20% improvement).

S → T	UCF → HMDB		HMDB → UCF	
Temporal Module	TemPooling	TemRelation	TemPooling	TemRelation
Target only	80.56 (-)	82.78 (-)	92.12 (-)	94.92 (-)
Source only	70.28 (-)	71.67 (-)	74.96 (-)	73.91 (-)
$\hat{G}_{sd}$	71.11 (0.83)	74.44 (2.77)	75.13 (0.17)	74.44 (1.05)
$\hat{G}_{td}$	71.11 (0.83)	74.72 (3.05)	75.13 (0.17)	75.83 (1.93)
$\hat{G}_{rd}$	- (-)	76.11 (4.44)	- (-)	75.13 (1.23)
All $\hat{G}_d$	71.11 (0.83)	**77.22 (5.55)**	75.13 (0.17)	**80.56 (6.66)**

Table 6: The full evaluation of accuracy (%) for integrating $\hat{G}_d$ in different positions without the attention mechanism. Gain values are in ().

S → T	UCF → HMDB		HMDB → UCF	
Temporal Module	TemPooling	TemRelation	TemPooling	TemRelation
Target only	80.56 (-)	82.78 (-)	92.12 (-)	94.92 (-)
Source only	70.28 (-)	71.67 (-)	74.96 (-)	73.91 (-)
All $\hat{G}_d$	71.11 (0.83)	77.22 (5.55)	75.13 (0.17)	80.56 (6.66)
All $\hat{G}_d$ +Domain Attn.	**73.06 (2.78)**	**78.33 (6.66)**	**78.46 (3.50)**	**81.79 (7.88)**

Table 7: The affect of the domain attention mechanism.

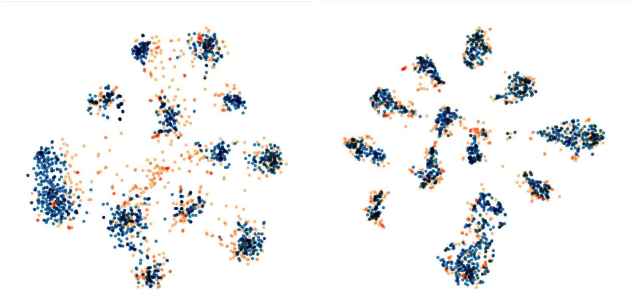
S → T	UCF → HMDB	HMDB → UCF
Target only	82.78 (-)	94.92 (-)
Source only	71.67 (-)	73.91 (-)
No Attention	77.22 (5.55)	80.56 (6.66)
General Attention	77.22 (5.55)	80.91 (7.00)
Domain Attention	**78.33 (6.66)**	**81.79 (7.88)**

Table 8: The comparison of different attention methods.

Attention mechanism. In addition to TemRelation, we also apply the domain attention mechanism to TemPooling by attending to the raw frame features instead of relation features, and improve the performance as well, as shown in Table 7. This implies that video DA can benefit from the domain attention even if the backbone architecture does not encode temporal dynamics. We also compare the domain attention module with the general attention module, which calculates the attention weights via the *FC-Tanh-FC-Softmax* architecture. However, it performs worse since the weights are computed within one domain, lacking of the consideration of domain discrepancy, as shown in Table 8.

Visualization of distribution. To investigate how our approaches bridge the gap between source and target domains, we visualize the distribution of both domains using t-SNE [28]. Figure 5 shows that TA^{3}N can group source data (blue dots) into denser clusters and generalize the distribution into the target domains (orange dots) as well.

Domain discrepancy measure. To measure the alignment between different domains, we use Maximum Mean Discrepancy (MMD) and domain loss, which are calculated using the final video representations. Lower MMD values and higher domain loss both imply smaller domain gap. TA^{3}N reaches lower discrepancy loss (0.0842) compared to the TemPooling baseline (0.184), and shows great improvement in terms of the domain loss (from 1.116 to 1.9286), as shown in Table 9.



(a) TemPooling + DANN [10] (b) TA^{3}N

Figure 5: The comparison of t-SNE visualization. The blue dots represent source data while the orange dots represent target data. See the supplementary for more comparison.

	Discrepancy loss	Domain loss	Validation accuracy
TemPooling	0.1840	1.1163	70.28
TemPooling + DANN [10]	0.1604	1.2023	71.11
TemRelation	0.2626	1.7588	71.67
TA^{3}N	**0.0842**	**1.9286**	**78.33**

Table 9: The discrepancy loss (MMD), domain loss and validation accuracy of our baselines and proposed approaches.

6. Conclusion and Future Work

In this paper, we present two large-scale datasets for video domain adaptation, **UCF-HMDB$_{full}$** and **Kinetics-Gameplay**, including both real and virtual domains. We use these datasets to investigate the domain shift problem across videos, and show that simultaneously aligning and learning temporal dynamics achieves effective alignment without the need for sophisticated DA methods. Finally, we propose **Temporal Attentive Adversarial Adaptation Network (TA^{3}N)** to simultaneously attend, align and learn temporal dynamics across domains, achieving state-of-the-art performance on all of the cross-domain video datasets investigated. The code and data are released here.

The ultimate goal of our research is to solve real-world problems. Therefore, in addition to integrating more DA approaches into our video DA pipelines, there are two main directions we would like to pursue for future work: 1) apply TA^{3}N to different cross-domain video tasks, including video captioning, segmentation, and detection; 2) we would like to extend these methods to the open-set setting [1, 35, 30, 13], which has different categories between source and target domains. The open-set setting is much more challenging but closer to real-world scenarios.

References

[1] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[2] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2, 5

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[4] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In *Domain Adaptation in Computer Vision Applications*, pages 1–35. Springer, 2017. 1, 2

[5] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[6] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)*, 2014. 2

[7] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[8] Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015. 1, 2, 3

[10] 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. 1, 2, 3, 6, 7, 8

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[13] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations (ICLR)*, 2018. 8

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015. 3

[15] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *British Machine Vision Conference (BMVC)*, 2018. 1, 3, 5, 6

[16] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2, 5

[18] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 1, 5

[19] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[20] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 1, 2, 6, 7

[21] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *International Conference on Learning Representations Workshop (ICLRW)*, 2017. 1, 2

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015. 1, 2

[23] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

[24] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning (ICML)*, 2017. 1, 2, 6, 7

[25] Xiang Long, Chuang Gan, Gerard de Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. Attention clusters: Purely attention based local feature integration for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[26] Chih-Yao Ma, Min-Hung Chen, Zsolt Kira, and Ghassan AlRegib. Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication*, 2018. 2

[27] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[28] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *The Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 8

[29] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359, 2010. 1

[30] Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2real: A new benchmark for synthetic-to-real visual domain adaptation. *arXiv preprint arXiv:1806.09755*, 2018. 1, 5, 6, 8

[31] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[32] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009. 1

[33] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[34] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3, 6, 7

[35] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *European Conference on Computer Vision (ECCV)*, 2018. 8

[36] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 4

[37] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2

[38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 5, 6

[39] Waqas Sultani and Imran Saleemi. Human action recognition across datasets by foreground-weighted histogram decomposition. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 3, 5, 6

[40] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshop (ECCVW)*, 2016. 2

[41] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 3

[42] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[43] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[44] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[45] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 3, 5

[46] Tiantian Xu, Fan Zhu, Edward K Wong, and Yi Fang. Dual many-to-one-encoder-based transfer learning for cross-dataset human action recognition. *Image and Vision Computing*, 55:127–137, 2016. 1, 3, 5, 6

[47] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[48] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[49] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[50] Xiao-Yu Zhang, Haichao Shi, Changsheng Li, Kai Zheng, Xiaobin Zhu, and Lixin Duan. Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 3

[51] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 4