This WACV 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.

Transferable-guided Attention Is All You Need for Video Domain Adaptation

André Sacilotti[1] Samuel Felipe dos Santos[2] Nicu Sebe[3] Jurandy Almeida[2]

[1]University of São Paulo
andre.sacilotti@usp.br

[2]Federal University of São Carlos
{samuel.felipe,jurandy.almeida}@ufscar.br

[3]University of Trento
niculae.sebe@unitn.it

Abstract

Unsupervised domain adaptation (UDA) in videos is a challenging task that remains not well explored compared to image-based UDA techniques. Although vision transformers (ViT) achieve state-of-the-art performance in many computer vision tasks, their use in video UDA has been little explored. Our key idea is to use transformer layers as a feature encoder and incorporate spatial and temporal transferability relationships into the attention mechanism. A Transferable-guided Attention (TransferAttn) framework is then developed to exploit the capacity of the transformer to adapt cross-domain knowledge across different backbones. To improve the transferability of ViT, we introduce a novel and effective module, named Domain Transferable-guided Attention Block (DTAB). DTAB compels ViT to focus on the spatio-temporal transferability relationship among video frames by changing the self-attention mechanism to a transferability attention mechanism. Extensive experiments were conducted on UCF-HMDB, Kinetics-Gameplay, and Kinetics-NEC Drone datasets, with different backbones, like ResNet101, I3D, and STAM, to verify the effectiveness of TransferAttn compared with state-of-the-art approaches. Also, we demonstrate that DTAB yields performance gains when applied to other state-of-the-art transformer-based UDA methods from both video and image domains. Our code is available at https://github.com/Andre-Sacilotti/transferattn-project-code.

1. Introduction

With the popularization of social media platforms focused on user-generated content, a huge volume of data is generated, for instance, 720,000 of hours of video content is uploaded to YouTube daily[1]. The cataloging and searching of this content is necessary, however, manually analyzing this immense amount of content is practically impossible, making video analysis tasks crucial.

[1]https://www.demandsage.com/youtube-stats/

Among the several video analysis tasks, action recognition is one of the most popular and challenging ones since there is a significant number of variations in the manner the action can be carried out and captured, for example, speed, duration, camera, and actor movement, occlusion, etc [8].

Various deep learning methods for action recognition are available in the literature. These approaches can be classified based on how they handle the temporal dimension. Some use 3D models to capture spatial and temporal features, while others treat spatial and temporal data separately or employ Recurrent Neural Networks (RNNs) to model the temporal dynamics [25]. Despite all the advances, the temporal structure of videos still poses some challenges for training deep learning models [18]. Human costs are high, as many video annotations are needed to yield good results. Obtaining and annotating a desirable amount of data is difficult for many application domains, requiring significant human effort and specific knowledge [43].

Unsupervised Domain Adaptation (UDA) can be used to reduce the cost of manually annotating data. In these strategies, the model is trained with labeled data from a source domain and unlabeled data from a target domain to perform well on the target domain's test set. Since there is a domain change between source and target, UDA methods must deal with the distribution mismatch generated by the domain gap, since the domains might have different backgrounds, illumination, camera position, etc [6]. Several methods have been proposed to address this issue, e.g., adversarial-based methods [12, 13, 42], metric-based methods [14, 31], and more recently, transformer-based methods [48, 50], achieving remarkable results. However, most of these works are for image UDA, and video UDA is considerably less explored and significantly more challenging, as it requires handling the temporal aspects of the data [8].

Only a few recent works [5, 6, 8, 9, 19, 28, 32, 41, 46, 52] tackle video UDA for action recognition using deep learning with strategies like contrastive learning, cross-domain attention mechanisms, self-supervised learning, and multi-modalities of data. An amount even lower of works [8, 19] explore Vision Transformer (ViT) architectures.

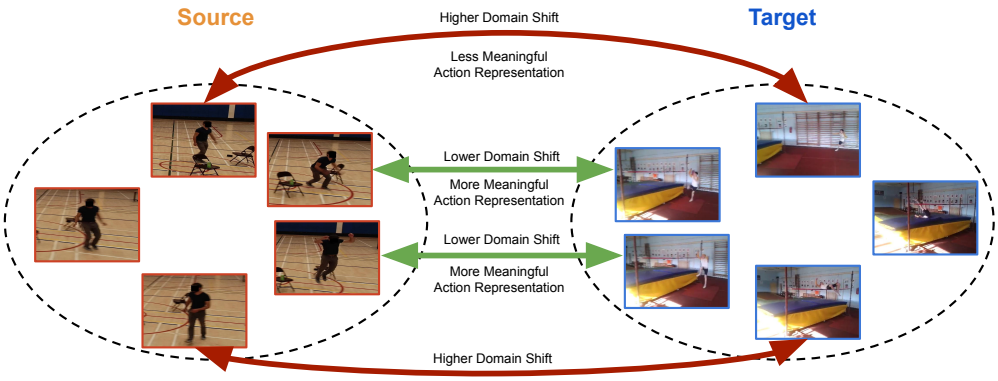


Figure 1. Main intuition behind our method. In this toy example, some frames in the *Jumping* class have more meaningful representation about the action, like the pose about to start the jump movement. Also, due to the capability to represent the important steps of the action, such frames can have minor domain shift compared to the rest of the video.

Although existing methods have improved the performance of UDA for action recognition, some limitations remains unexplored. First, they focus on aligning frames with a larger spatial and temporal domain gap across an entire video, but, as shown in Figure 1, we hypothesize that frames with a smaller domain gap have a more meaningful action representation and can improve the adaptability. Also, none of them investigate how to improve ViT for video UDA.

To address the mentioned issues, we propose a novel method for video UDA, called Transferable-guided Attention (TransferAttn), which improves the potential of transformer architecture. Our method uses pre-trained frozen backbones to extract frame-by-frame features from videos. A transformer encoder is used to reduce the domain gap and learn temporal relationships among frames. The encoder also includes our proposed transformer block, named Domain Transferable-guided Attention Block (DTAB), which introduces a new attention mechanism.

We evaluate our approach on three well-known video UDA benchmarks for action recognition, namely UCF $\leftrightarrow$ HDMB$_{full}$ [5], Kinetics $\rightarrow$ Gameplay [5], and Kinetics $\rightarrow$ NEC-Drone [7], where we outperform the other state-of-the-art methods. We also integrated our proposed DTAB module into other state-of-the-art transformer architectures for UDA, showing that it may increase their performance.

The main contributions of this paper are the following:

- To the best of our knowledge, we are the first to present a backbone-independent transformer architecture for video UDA. Our experiments showed the effectiveness of the transformer encoder in extracting fine-grained spatio-temporal transferable representations.
- We propose DTAB, a novel transferable transformer block for UDA. Our method employs a new attention mechanism that improves adaptation and domain transferability. Also, we show the positive effect of applying the DTAB module to other state-of-the-art UDA methods, both for videos and images.
- We conduct extensive experiments on several benchmarks, setting a new state-of-the-art result in three different cross-domain datasets. Also, our ablation study shows the positive effect of each part of our approach.

2. Related Work

2.1. Video-based Action Recognition

Action recognition methods have been extensively studied with the advent of deep learning, especially with the introduction of large-scale video datasets, such as Kinetics, Moments-In-Time, YouTube Sports 1M, and Youtube 8M [20]. BEAR [10] states a new benchmark in action recognition, which is made to cover a diverse set of real-world applications. Deep learning-based solutions can be divided into three categories according to how they model the temporal dimension [25]: (1) space-time networks, (2) multi-stream networks, and (3) hybrid models. Space-time networks use 3D convolutions to maintain temporal information, inflating 2D kernels to 3D, like C3D [40] and I3D [4]. Multi-stream networks employ different models to deal with spatial (usually RGB images) and temporal (usually optical flow) information, like TSN [45], which applies temporal sampling; and TDN [44], which has modules to capture short-term and long-term motion (across segments). Hybrid models integrate recurrent networks, like LSTMs [11, 47, 53] and Temporal CNNs [22], on top of CNNs. Skeleton data, like body joint information, can also be utilized [22, 29, 35, 54] and recent works [37, 49] show that graph convolution obtains superior performance to RNNs and Temporal CNNs on capturing information from joints [25]. Kim et al. [24] presents a novelty training approach to make models robust to distribution shifts.

2.2. Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) in the image domain has a wide range of strategies to address the domain shift. A standard option is the adversarial-based methods [12, 13, 27, 42], which use a domain discriminator while maximizing the feature extractor loss through a min-max optimization game, similar to Generative Adversarial Networks (GAN) [16] training, minimizing the domain gap. In addition, the metric-based methods aim to reduce the domain gap by learning domain-invariant features through discrepancy metrics, like Maximum Mean Discrepancy (MMD) [14] and Joint Adaptation Networks (JAN) [31], that incorporate a loss metric computing the discrepancy between the domain features and aim to reduce that metric to minimize the domain shift. Driven by the success of ViTs, CDTrans [48] adopts a three-branch cross transformer that proves to be noisy-robust. On the other side, TVT [50] employs a transferability metric as a weight into class token attention weight. Although TVT [50] shows great results injecting transferability into the class token weight, it lacks two essential points: i) TVT [50] does not use spatial relation transferability; ii) As an image UDA, it does not incorporate the temporal relation transferability.

2.3. UDA for Action Recognition

Although there are several possible applications of UDA for action recognition in real-world problems, only a limited number of recent studies have tackled this challenging task [5, 6, 9, 28, 32, 41, 46, 52]. TA^{3}N [5] proposes a domain attention mechanism that focuses on the temporal dynamics of the videos. MA^{2}LT-D [6] generates multi-level temporal features with multiple domain discriminators. Level-wise attention weights are calculated by domain confusion and features are aggregated by attention determined by the domain discriminators. Other approaches use multiple modalities of data, like MM-SADA [32], where self-supervision among modalities is used, and MixDANN [52], which dynamically estimates the most adaptable modality and uses it as a teacher to the others. CleanAdapt [9] tackles the source-free video domain adaptation problem using a model pre-trained on the source domain to generate noisy labels for the target domain, and then the likely correct ones are used to fine-tune the model. STHC [28] addresses the source-free domain using spatial and temporal augmentation. In a different direction, TranSVAE [46] handles spatial and temporal domain divergence separately by constraining different sets of latent factors.

Although transformers can obtain state-of-the-art performance, only a few works for video UDA exist. UDAVT [8] is a recent work that leverages the STAM visual transformer [36] and proposes a domain alignment loss based on the Information Bottleneck (IB) principle to learn domain invariant features. Also, MTRAN [19], which depends on 3D backbones, uses a transformer layer inspired by ViViT [1], where each token is a 16-frame clip representation. Although UDAVT and MTRAN show great results incorporating the transformer mechanism, the UDAVT architecture strictly depends on transformer backbones that deal separately with spatial and temporal relations, like STAM [36]. At the same time, MTRAN is dependent on 3D backbones, and the attention relation is done on clip-level pooled features, lacking a more fine-grained frame-level relation. Also, none of them exploit ways to improve knowledge transferring in the transformer mechanism.

3. Our Approach

First, in Section 3.1, we introduce our baseline model, which uses ViT as a feature encoder for a simple adversarial domain adaptation approach. Then, in Section 3.2, we introduce our domain transferable-guided attention block, called DTAB, and its components.

3.1. Baseline Model

Given the success of ViTs in diverse computer vision tasks [2, 15, 30], we built a baseline architecture that leverages a ViT encoder to produce better spatio-temporal features for action recognition and an adversarial-based method for UDA, as shown in Figure 2.

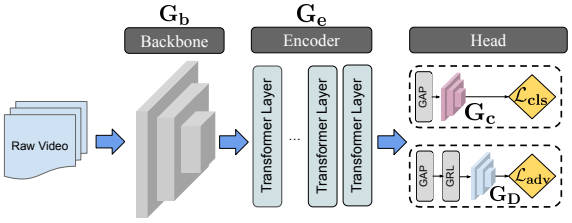


Figure 2. A baseline architecture trained with an adversarial loss $\mathcal{L}_{adv}$ and a classification loss $\mathcal{L}_{cls}$. The backbone G_b extracts frame-level features and the encoder G_e learns meaningful semantic spatio-temporal representations.

Overall, the baseline architecture consists of a backbone, an encoder, a classification head, and an adaptation head. The backbone (G_b) is fixed and not trained. The encoder (G_e) comprises L transformer layers, with h attention heads and a hidden size of d. Before the encoder G_e, a patch embedding (G_p) may be used to map the features produced by the backbone G_b to the input size of the encoder G_e.

Let $\mathcal{S}$ be a labeled source domain sharing the same set $\mathcal{Y}$ of categories with an unlabeled target domain $\mathcal{T}$, where $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ is a set of tuples composed of N_s videos x_i^s and their respective label y_i^s. Since we do not know the labels of the target domain beforehand, the set $\mathcal{T} = \{x_i^t\}_{i=1}^{N_t}$ comprises the target videos only. Once k frames are uniformly sampled from each video, we denote the j-th frame

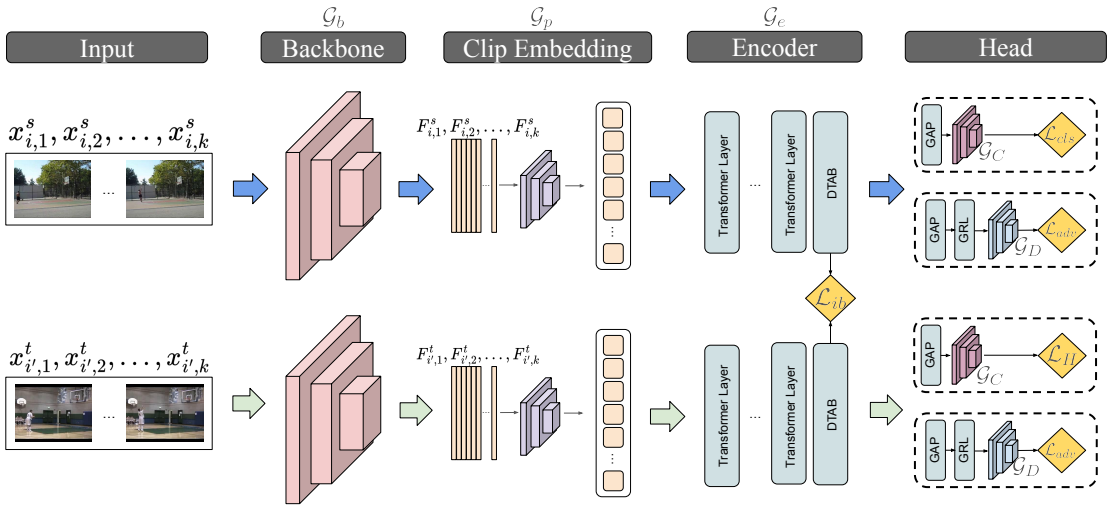


Figure 3. Overview of our TransferAttn. The video frames are fed into a fixed backbone to extract frame-by-frame features, followed by a clip embedding to map frames into tokens. The embeddings are fed into a sequence of transformers to extract transferable spatio-temporal information. The adaptation branch for adversarial domain discrimination uses fine-grained representations from the transformer encoder.

from the i-th video as $x^s_{i,j}$ for the source domain and $x^t_{i,j}$ for the target domain. For convenience, we refer to the features extracted from the backbone G_b for the i-th video and the j-th frame from the source domain as $F^s_{i,j}$, and $F^t_{i,j}$ for the target domain. Also, we denote f^s_i and f^t_i as a Global Average Pooling (GAP) over the frame-level features extracted by the encoder G_e for the i-th video from the source and target domains, respectively.

The classification head contains a classifier (G_C). Unlike previous works, where the classifier G_C is trainable, in our framework, the classifier G_C is not trained. Therefore, instead of learning a mapping from features to classes, our model uses a MLP classifier with fixed random weights. A similar idea is also adopted by UDAVT [8] in the projection head used to map source and target domain features. By using a fixed classifier, the classification boundaries do no change during training and the encoder G_e is enforced to learn a feature space that better fits to the class separability imposed by the classifier G_C. The motivation for using a fixed classifier is to prevent it from learning a projection that might overfit the source domain data. This way, we make learning class discrimination a responsibility of the encoder G_e. Although the classification branch is not trained, the classification ($\mathcal{L}_{cls}$) and entropy ($\mathcal{L}_H$) losses, as given by Equations 1 and 2, respectively, are used to train the model, ensuring that the encoder G_e learns a feature space from both source and target domain data. As far as we know, we are the first to propose this idea for video UDA.

$$\mathcal{L}_{cls} = -\frac{1}{N_s}\sum_{i=1}^{N_s} y^s_i \cdot \log G_C(f^s_i) \tag{1}$$

$$\mathcal{L}_H = -\frac{1}{N_t}\sum_{i=1}^{N_t} G_C(f^t_i) \cdot \log G_C(f^t_i) \tag{2}$$

The adaptation branch is composed of a Gradient Reversal Layer (GRL) followed by a domain discriminator (G_D), which is a MLP classifier trained to identify if a video is from the source or target domain. At the same time, as the gradients are inverted in the GRL, the encoder G_e is trained to deceive the domain discriminator G_D, playing a min-max game. This is achieved with the adversarial domain discriminator loss ($\mathcal{L}_{adv}$), as in Equation 3.

$$\mathcal{L}_{adv} = -\frac{1}{N_s}\sum_{i=1}^{N_s} \log G_D(\text{GRL}(f^s_i)) - \frac{1}{N_t}\sum_{i=1}^{N_t} \log G_D(\text{GRL}(f^t_i)) \tag{3}$$

Therefore, the overall loss used to train the baseline model is a weighted sum of three terms: the classification loss $\mathcal{L}^s_{cls}$ (Equation 1), the entropy loss $\mathcal{L}^t_H$ (Equation 2), and the adversarial loss $\mathcal{L}_{adv}$ (Equation 3).

3.2. TransferAttn: Transferable-guided Attention

The baseline architecture with standard ViT encoder, as shown in Tables 1, 2, 3, achieves limited improvement for video UDA in the action recognition task. Motivated by this observation, we hypothesize that merely using the standard self-attention ViT is insufficient to fully leverage the transferable capabilities of ViT, as it overlooks two crucial factors: (*i*) it does not consider inter-domain specificity and (*ii*) it focus only on intra-domain frames similarity.

To support our hypothesis, we propose TransferAttn, an improvement of our baseline model, as shown in Figure 3. The key difference between TransferAttn and our

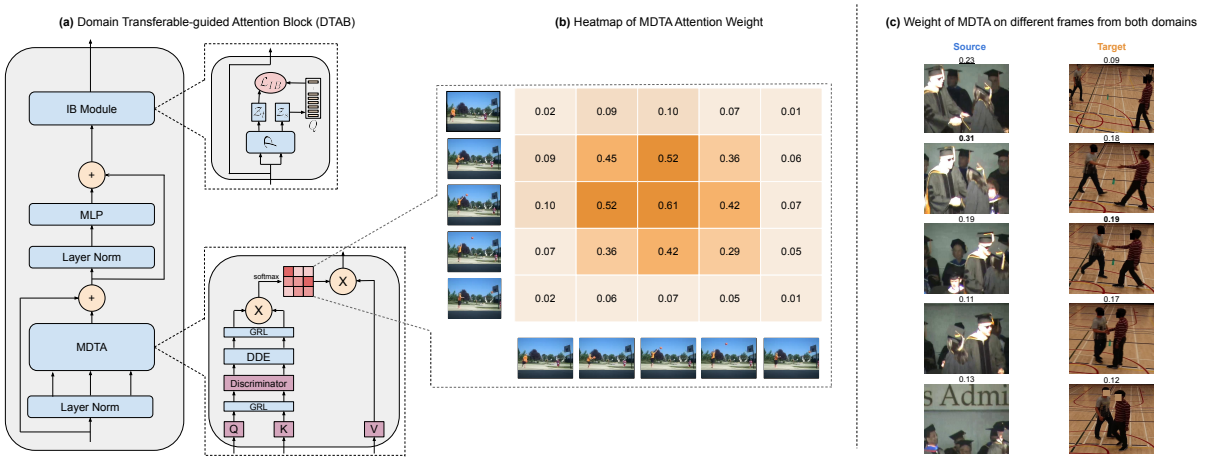


Figure 4. DTAB overview. (a) DTAB follows a standard transformer block, except for our novel MDTA mechanism and the layer-wise IB calculation. (b) Heatmap of transferable-attention weights, showing how MDTA focuses on frames that are more transferable between domains and also brings more meaningful information about the action. (c) Temporal attention visualization compared between domains.

baseline model is in the encoder G_e. To ensure the encoder G_e is capable of learning a better feature space, we design a novel transformer module to facilitate the domain alignment, named Domain Transferable-guided Attention Block (DTAB), as shown in Figure 4. Roughly, DTAB is a transformer block where the self-attention layer is replaced by a new attention mechanism and a domain alignment module is added before the output (see Figure 4(a)).

Instead of using self-attention, which gives attention to similar tokens or patches (in our case, frames) only within a same domain context, we propose the Multi-head Domain Transferable-guided Attention (MDTA), a transferable-guided attention that acts as a cross-domain attention. This mechanism weights frames based on how difficult it is to predict which domain they belong to (see Figure 4(b)), pushing the model to focus on local temporal features that have a small domain gap. Unlike methods that prioritize frames with larger domain gaps, like TA3N [5] and MA2LT-D [6], the reasoning behind MDTA is to highlight portions of a video that are more transferable, or in other words, similar across different domains. As an example, Figure 4(c) shows the weights produced by MDTA for each frame of a video taken from source and target domains. Note that MDTA enables the model to focus on more significant frames in the action *Hand Shaking* that are more transferable between different domains.

MDTA does not align the source and target domains, but only keeps the attention on portions of a video that are likely to deceive a discriminator in predicting which domain they come from. To perform domain alignment, DTAB exploits the Information Bottleneck (IB) principle [23, 33]. The IB mechanism adopted in DTAB was introduced by UDAVT [8], but unlike UDAVT, DTAB applies IB to local features. The IB mechanism helps to find a shared representation space where the weighted local features from different domains coming out from MDTA are mapped so that the shared information between domains is maximized. DTAB works not only on our TransferAttn, but also improves the results of other state-of-the-art methods, both for video and image UDA, as reported in Tables 4 and 5. As far we know, we are the first to propose this idea for video UDA.

Multi-head Domain Transferable-guided Attention. To overpass the main bottlenecks related to the standard self-attention mechanism, we propose the Multi-head Domain Transferable-guided Attention (MDTA), an new attention mechanism that leverages spatio-temporal information to give more weight to frames or patches that are more similar across domains and still semantically meaningful. Similar to self-attention, MDTA is computed as in Equations 4 and 5,

$$\mathrm{MDTA}(Q,K,V) = \mathrm{Concat}(\{\mathrm{DTA}_i\}_{i=1}^{h}) \cdot W_O, \quad (4)$$

$$\mathrm{DTA}_i = \mathrm{softmax}\left(\frac{Q^{'} \cdot K^{'T}}{\sqrt{d_h}}\right) \cdot V \cdot W_i^V, \quad (5)$$

where W_i^Q, W_i^K, and W_i^V are linear projections learned separately for each of the h heads; and W^O is the learned linear projection of the concatenation.

Unlike self-attention, in MDTA, the attention scores are computed by taking the dot product between vectors $Q^{'} = \mathrm{GRL}(\mathrm{DDE}(G_{D'}(\mathrm{GRL}(QW_i^Q))))$ and $K^{'} = \mathrm{GRL}(\mathrm{DDE}(G_{D'}(\mathrm{GRL}(KW_i^K))))$, where the Domain Discriminator Error (DDE) is a weighting function given by Equation 6 and $G_{D'}$ is a domain discriminator that predicts which domain the frames of a video belong to. The purpose of the double GRL is to force the domain discriminator $G_{D'}$ to learn domain-separable class-invariant features.

$$\mathrm{DDE}(x) = \begin{cases} \log x, & \text{if } x \text{ is from source} \\ \log(1-x), & \text{otherwise} \end{cases} \quad (6)$$

This operation results in a matrix that defines which frames are more or less transferable, as depicted in Figure 4(b), where the higher transferability values are given to the middle frames, probably because they are more meaningful and where the action actually takes place and are therefore likely similar to other domains. In other words, if DDE goes to one, it is more likely to deceive the domain discriminator $G_{D'}$ because its features are not easy to predict which domain they come from and should be more important when classifying the video action.

Information Bottleneck Module. Since MDTA does not align the source and target domains, we propose to integrate it with a Information Bottleneck (IB) module [8]. To perform domain alignment, this module calculates a loss function given by Equation 7,

$$\mathcal{L}_{ib} = \sum_{i=1}^{m}(1 - C_{i,i})^2 + \lambda \sum_{i=1}^{m}\sum_{j \neq i}^{m}(C_{i,j})^2 \quad (7)$$

where C is a cross-correlation matrix whose elements $C_{i,j} = \frac{\sum_b^B z_{i,b}^s \cdot z_{j,b}^t}{\sqrt{\sum_b^B (z_{i,b}^s)^2}\sqrt{\sum_b^B (z_{j,b}^t)^2}}$ are computed over mean-centered representations z_i^s and z_j^t of the source and target domains, respectively, taken from a batch of B out of a total of m features. We also utilize a queue, $\mathcal{Q}$, to store source representations, thereby increasing the number of pairs used in the correlation calculations for each batch.

The loss function used to train our TransferAttn model is a weighted sum of four terms: classification loss ($\mathcal{L}_{cls}^s$), entropy loss ($\mathcal{L}_H^t$), adversarial loss ($\mathcal{L}_{adv}$), and IB loss ($\mathcal{L}_{ib}$).

4. Experiments and Results

To assess the effectiveness of our approach, we carried out extensive experiments on several datasets for video UDA and compared against state-of-the-art methods.

4.1. Datasets

Experiments were conducted on 3 different benchmarks:

UCF $\leftrightarrow$ HDMB$_{full}$ [5] is one of the most widely used, containing a subset of videos from two public datasets, UCF101 [39] and HMDB51 [26], representing a total of 3209 videos and 12 classes.

Kinetics $\rightarrow$ Gameplay [5] is a non-public dataset that contains a subset of videos from the well-known Kinetics-400 [21] and a private gameplay dataset [5], containing 49998 videos and 30 classes.

Kinetics $\rightarrow$ NEC-Drone [7] is a public dataset that contains videos from Kinetics-600 [3] and NEC-Drone. The dataset contains 10118 videos and 7 classes. For a fair comparison, we used the cropped version of NEC-Drone [8].

4.2. Experimental Setup

For extracting features from the videos, we tested three different backbones, ResNet101 [17], I3D [4], and STAM [36], all of them pretrained. Our approach relies on frame-level features, so for the 3D backbones, like STAM [36] and I3D [4], we densely slide a temporal window of 16 frames along each video.

Our encoder G_e comprises 4 transformer blocks, where our DTAB module is the last one, and each transformer block is composed of $h = 8$ attention heads with a hidden size of $d_{model} = 512$. We used the ADAM optimizer for the training schedule with a weight decay of $5 \cdot 10^{-4}$ and a learning rate of $3 \cdot 10^{-5}$ for 300 epochs. We present the adversarial training, DTAB hyperparameters and sampling parameters in the supplementary material.

4.3. Comparison with State-of-the-art Methods

In this section, we compare our approach with different methods recently proposed in the literature.

4.3.1 Results on UCF101$\leftrightarrow$HMDB51$_{\text{full}}$

As shown in Table 1, we compare our results using three different backbones. The first one is ResNet101, a 2D backbone, in which our approach achieves a significant average increase of 2.4%, resulting in 88.2%. This accuracy surpasses some methods that use the I3D backbone, showing that even with a 2D backbone, our approach can extract significant spatio-temporal information.

By analyzing methods that use the I3D backbone, our approach surpasses even works that use multi-modal data (i.e., color and motion). From single-modal data, our approach yields a significant average increase of 3.5% and, compared to multi-modal methods, an average increase of 0.4%.

Finally, for the STAM backbone, it can be seen in Table 1 that our approach yields a significant average increase of 0.9%. Also, our TransferAttn model achieves a significant improvement compared to our baseline model.

4.3.2 Results on Kinetics$\rightarrow$Gameplay

This section presents the results of our approach for the Kinetics$\rightarrow$ Gameplay benchmark. Since the Gameplay dataset does not give access to the raw videos, but only to frame-level features extracted from ResNet101, we were unable to provide results for methods where the backbone is not fixed in the training stage or relies on video-level features. As shown in Table 2, our approach achieves a significant increase of 6.5% in accuracy compared with MA^2LTD, a state-of-the-art method, and is better than all previous baselines. Also, it can be seen that our TransferAttn model performs significantly better than our baseline model.

Table 1. Classification Accuracy on UCF101↔ HMDB51$_{\text{full}}$. Multi-modal methods are represented with (C + M).

Method	Backbone	U → H	H → U	Average
Source Only	ResNet101	73.9	71.7	72.8
TA^{3}N [5]		78.3	81.8	80.1
MA^{2}LT-D [6]		85.0	86.6	85.8
Baseline		84.2	85.4	84.8
TransferAttn (ours)		**88.1**	**88.3**	**88.2**
Source Only	TRN	82.2	88.1	85.2
STHC [28]		90.9	92.1	91.5
TranSVAE [46]		92.2	96.5	94.3
Baseline		91.2	93.6	92.4
TransferAttn (ours)		**93.5**	**97.1**	**95.3**
Source Only	I3D	80.6	89.3	85.0
STCDA (C + M) [38]		83.1	92.1	87.7
MA^{2}LT-D [6]		89.3	91.2	90.3
CO^{2}A [41]		87.8	95.8	91.8
CIA (C + M) [51]		91.9	94.6	93.3
TranSVAE [46]		87.8	99.0	93.4
MTRAN (C + M) [19]		92.2	95.3	93.8
CleanAdapt (C + M) [9]		93.6	99.3	96.5
Baseline		90.8	95.2	93.0
TransferAttn (ours)		**94.4**	**99.4**	**96.9**
Source Only	STAM	86.9	93.7	90.3
TranSVAE [46]		93.5	99.5	96.5
UDAVT [8]		92.3	96.8	94.6
MA^{2}LT-D [6]		95.3	99.4	97.4
Baseline		93.4	98.9	96.1
TransferAttn (ours)		**97.2**	**99.7**	**98.5**

Table 2. Classification Accuracy on Kinetics → Gameplay.

Method	Backbone	K → G
Source Only	ResNet101	17.6
TA^{3}N [5]		27.5
MA^{2}LT-D [6]		31.5
TranSVAE [46]		21.9
Baseline		30.4
TransferAttn (ours)		**37.0**

4.3.3 Results on Kinetics→NEC-Drone

Our approach was also evaluated in the Kinetics → NEC-Drone benchmark, as we can see in Table 3, achieving a significant increase of 9.5% in comparison with UDAVT, establishing a new state-of-the-art result.

Table 3. Classification Accuracy on Kinetics → NEC-Drone.

Method	Backbone	K → N
Source Only	STAM	29.4
MA^{2}LT-D [6]		55.4
TranSVAE [46]		55.9
UDAVT [8]		65.3
Baseline		45.5
TransferAttn (ours)		**74.8**

Table 4. Accuracy in HMDB-UCF$_{full}$ dataset integrating DTAB on state-of-the-art transformer-based video UDA architectures.

Method	U → H	H → U	Average
UDAVT (Our impl.) [8]	92.2	96.5	94.4
UDAVT + TAM [50]	92.5	96.9	94.7
UDAVT + DTAB	**94.2**	**97.9**	**96.1**

4.4. Ablation Study

This section presents an ablation study, showcasing how each design aspect of our approach impacts its performance, and shows how our DTAB can improve the results of other transformer-based methods for both action recognition (i.e., video UDA) and image classification (i.e., image UDA).

4.4.1 Effect of DTAB components

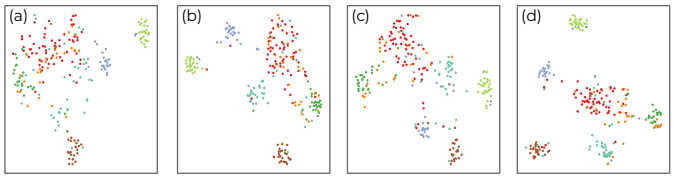
To understand the individual contributions from the components that integrate our DTAB module and how they improve its domain transferability, we conducted an in-depth study to evaluate how each component of DTAB impacts its performance. Figure 5 summarizes the obtained results, showing the accuracy for the Kinetics → NEC-Drone benchmark and the t-SNE visualization. By replacing the self-attention mechanism with our MDTA in the last transformer block, the accuracy is increased by 9.3%. On the other hand, using the IB mechanism in the last transformer block, even with the standard self-attention mechanism, increases the accuracy by 12.8%. Finally, integrating all the components, we achieve a significant increase of 29.3%, indicating the importance of both components in reducing the domain gap. Also, the t-SNE plots clearly show how each component impacts the clusterization of the low-dimensional features from the different classes, and it is easy to see that the combination of all the components makes features from a same class better clustered while different classes are farther away.

4.4.2 Effects of DTAB on other methods for video UDA

Another aspect we studied is the effect of our new transformer block, DTAB, on other transformer-based methods for video UDA. For this study, we first ran UDAVT [8] on the UCF→HMDB$_{\text{full}}$ benchmark and reported the results obtained by our reproduction using the author's code[2]. As a baseline, we also reported the results for the Transferability Adaption Module (TAM), which is the attention mechanism introduced in TVT [50]. Due to its source-free characteristics, we were unable to evaluate DTAB on MTRAN [19].

The results in Table 4 show that adding our DTAB increased the accuracy by 1.7%, while TAM yielded marginal

[2] https://github.com/vturrisi/UDAVT



#	Components	K → N
a)	Baseline	45.5
b)	+ IB	54.8
c)	+ MDTA	58.3
d)	DTAB (MDTA + IB)	74.8

Figure 5. Ablation study on Kinetics → NEC-Drone integrating each component of DTAB separately in comparison with standard transformer. **Left:** The t-SNE plots for class-wise features. **Right:** The accuracy result of each component.

Table 5. Classification Accuracy in Office-31 dataset integrating DTAB on state-of-the-art transformer-based image UDA architectures.

Method	Backbone	A → W	D → W	A → D	D → A	W → A	Avg.
TVT (Our impl.) [50]	ViT-B 16	93.6	98.7	93.7	79.4	78.7	88.8
TVT + DTAB		94.7	99.9	95.0	80.3	79.5	**89.9**
CDTrans (Our impl.) [48]	DeiT-S	93.5	98.2	94.0	77.7	77.0	88.1
CDTrans + DTAB		94.7	98.4	95.6	77.7	78.1	**88.9**

gains of less than 0.5%. These results demonstrate that DTAB outperforms TAM in handling spatio-temporal information. Also, they suggest that our DTAB module can integrate with other transformer-based methods for video UDA, enhancing their ability to reduce the domain gap.

4.4.3 Effects of DTAB on transformers for image UDA

We also tested the effects of using our DTAB module with TVT [50] and CDTrans [48], which are transformer-based methods devised for image UDA. For this study, experiments were conducted on the Office-31 dataset [34], which is composed of 4,652 images from 31 classes and comprises three different domains, namely: Amazon (A), DSRL (D), and Webcam (W). The results reported in Table 5 were obtained by our reproduction using the authors' code[3,4]. These results show a slight increase of 1.1% in the accuracy of TVT [50]; while for CDTrans [48], we observe an increase of 0.8% in accuracy.

4.4.4 Complexity Analysis

We also conduct a complexity analysis of our TransferAttn model. In this study, we compare TransferAttn and other baselines with respect to the amount of trainable parameters (#Parameters) and floating point operations (GFLOPs). As shown in Table 6, our approach has slightly more trainable parameters than TranSVAE [46], however, this difference is compensated by the significant improvement in UDA results reported in Section 4.3. Compared with other transformer-based methods, like UDAVT [8], our approach has much less trainable parameters and floating point operations, in addition to a significant improvement in UDA results, indicating the efficiency of our TransferAttn model.

[3] https://github.com/uta-smile/TVT/
[4] https://github.com/CDTrans/CDTrans/

Table 6. Comparison on Model Complexity

Methods	#Parameters	GFLOPs
TranSVAE	12.7 M	36.5
MA^2LT-D	49.7 M	51.2
UDAVT	118.9 M	524.7
TransferAttn (ours)	16.3 M	55.3

5. Conclusions

Only a few works have exploited transformer architectures for video UDA, a promising strategy due to their performance in other tasks. Motivated by this observation, we proposed TransferAttn, a framework for video UDA and one of the few works that exploit transformer architectures to adapt cross-domain knowledge. We also propose a novel Domain Transferable-guided Attention Block (DTAB) that employs an attention mechanism to encourage spatial-temporal transferability among video frames from different domains. We outperformed all other state-of-the-art methods we compared, showing the effectiveness of our TransferAttn model. Our DTAB module also demonstrated to be a promising strategy by itself and when added to other state-of-the-art UDA methods, it increased their performance.

As future work, we intend to evaluate our approach on other video UDA tasks, like video segmentation and action localization. In addition, we also plan to augment our approach to utilize multi-modal data and make it able to integrate into source-free methods.

Acknowledgments This research was supported by São Paulo Research Foundation - FAPESP (grants #2023/17577-0 and #2024/04500-2) and Brazilian National Council for Scientific and Technological Development - CNPq (grants #315220/2023-6 and #420442/2023-5).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In IEEE International Conference on Computer Vision (ICCV), 2021. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European Conference on Computer Vision (ECCV), pages 213–229. Springer, 2020. 3

[3] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. CoRR, abs/1808.01340, 2018. 6

[4] 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, 6

[5] Min-Hung Chen, Zsolt Kira, Ghassan Alregib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In IEEE International Conference on Computer Vision (ICCV), pages 6320–6329, 2019. 1, 2, 3, 5, 6, 7

[6] Peipeng Chen, Yuan Gao, and Andy J. Ma. Multi-level attentive adversarial learning with temporal dilation for unsupervised video domain adaptation. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 776–785, 2022. 1, 3, 5, 7

[7] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1706–1715, 2020. 2, 6

[8] V. da Costa, G. Zara, P. Rota, T. Oliveira-Santos, N. Sebe, V. Murino, and E. Ricci. Unsupervised domain adaptation for video transformers in action recognition. In IEEE International Conference on Pattern Recognition (ICPR), pages 1258–1265, 2022. 1, 3, 4, 5, 6, 7, 8

[9] Avijit Dasgupta, C.V. Jawahar, and Karteek Alahari. Overcoming label noise for source-free unsupervised video domain adaptation. In Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP). ACM, 2023. 1, 3, 7

[10] Andong Deng, Taojiannan Yang, and Chen Chen. A large-scale study of spatiotemporal representation learning with a new benchmark on action recognition. CoRR, abs/2303.13505, 2023. 2

[11] 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), pages 2625–2634, 2015. 2

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In International Conference on Machine Learning (ICML), page 1180–1189. PMLR, 2015. 1, 3

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks, pages 189–209. Springer International Publishing, 2017. 1, 3

[14] Muhammad Ghifary, W. Bastiaan Kleijn, and Mengjie Zhang. Domain adaptive neural networks for object recognition. In PRICAI 2014: Trends in Artificial Intelligence, pages 898–904. Springer International Publishing, 2014. 1, 3

[15] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 244–253, 2019. 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Conference on Neural Information Processing Systems (NeurIPS), volume 27. Curran Associates, Inc., 2014. 3

[17] 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. 6

[18] D-A. Huang, V. Ramanathan, D. Mahajan, L. Torresani, M. Paluri, L. Fei-Fei, and J. C. Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7366–7375, 2018. 1

[19] Yi Huang, Xiaoshan Yang, Ji Zhang, and Changsheng Xu. Relative alignment network for source-free multimodal video domain adaptation. In International Conference on Multimedia (MM), page 1652–1660. ACM, 2022. 1, 3, 7

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 35(1):221–231, 2013. 2

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. CoRR, abs/1705.06950, 2017. 6

[22] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3288–3297, 2017. 2

[23] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A. Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain Self-supervised Learning for Domain Adaptation with Few Source Labels. CoRR, abs/2003.08264, 2020. 5

[24] Kiyoon Kim, Shreyank N Gowda, Panagiotis Eustratiadis, Antreas Antoniou, and Robert B Fisher. Adversarial augmentation training makes action recognition models more robust to realistic video distribution shifts. CoRR, abs/2401.11406, 2024. 2

[25] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. International Journal of Computer Vision (IJCV), 130(5):1366–1401, 2022. 1, 2

[26] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In IEEE International Conference on Computer Vision (ICCV), pages 2556–2563, 2011. 6

[27] Zhengfeng Lai, Haoping Bai, Haotian Zhang, Xianzhi Du, Jiulong Shan, Yinfei Yang, Chen-Nee Chuah, and Meng Cao. Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 2691–2701, 2024. 3

[28] K. Li, D. Patel, E. Kruus, and M. Min. Source-free video domain adaptation with spatial-temporal-historical consistency learning. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 14643–14652, 2023. 1, 3, 7

[29] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In European Conference on Computer Vision (ECCV), pages 816–833. Springer, 2016. 2

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In IEEE International Conference on Computer Vision (ICCV), pages 10012–10022, 2021. 3

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In International Conference on Machine Learning (ICML), page 2208–2217. PMRL, 2017. 1, 3

[32] Jonathan Munro and Dima Damen. Multi-modal Domain Adaptation for Fine-grained Action Recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1, 3

[33] Changhwa Park, Jonghyun Lee, Jaeyoon Yoo, Minhoe Hur, and Sungroh Yoon. Joint Contrastive Learning for Unsupervised Domain Adaptation. CoRR, abs/2006.10297, 2020. 5

[34] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, European Conference on Computer Vision (ECCV), pages 213–226. Springer, 2010. 8

[35] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1010–1019, 2016. 2

[36] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? CoRR, abs/2103.13915, 2021. 3, 6

[37] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1227–1236, 2019. 2

[38] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 9787–9795, 2021. 7

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR, abs/1212.0402, 2012. 6

[40] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In IEEE International Conference on Computer Vision (ICCV), pages 4489–4497, 2015. 2

[41] Victor G. Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. Dual-head contrastive domain adaptation for video action recognition. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 2234–2243, 2022. 1, 3, 7

[42] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2962–2971, 2017. 1, 3

[43] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 27(12):2591–2600, 2017. 1

[44] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1895–1904, 2021. 2

[45] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 41(11):2740–2755, 2018. 2

[46] Pengfei Wei, Lingdong Kong, Xinghua Qu, Yi Ren, zhiqiang xu, Jing Jiang, and Xiang Yin. Unsupervised video domain adaptation for action recognition: A disentanglement perspective. In Conference on Neural Information Processing Systems (NeurIPS). Curran Associates, Inc., 2023. 1, 3, 7, 8

[47] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In International Conference on Multimedia (MM), pages 461–470. ACM, 2015. 2

[48] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. CoRR, abs/2109.06165, 2021. 1, 3, 8

[49] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In AAAI Conference on Artificial Intelligence, volume 32, 2018. 2

[50] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 520–530, 2023. 1, 3, 7, 8

[51] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for

domain adaptive action recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 14702–14712, 2022. 7

[52] Yuehao Yin, Bin Zhu, Jingjing Chen, Lechao Cheng, and Yu-Gang Jiang. Mix-dann and dynamic-modal-distillation for video domain adaptation. In International Conference on Multimedia (MM), page 3224–3233. ACM, 2022. 1, 3

[53] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4694–4702, 2015. 2

[54] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In AAAI Conference on Artificial Intelligence, volume 30, 2016. 2