TransferAttn: Transferable-guided Attention for Video Domain Adaptation

André Sacilotti
Inst. Math. & Comput. Sci.
University of São Paulo
Email: andre.sacilotti@usp.br

Nicu Sebe
Dept. Inf. Eng. & Comput. Sci.
University of Trento
Email: niculae.sebe@unitn.it

Jurandy Almeida
Dept. Computing
Federal University of São Carlos
Email: jurandy.almeida@ufscar.br

***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 domain adaptation has still been little explored. Our key idea is to use the 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), which compels ViT to focus on the spatio-temporal transferability relationship among video frames by changing the self-attention mechanism to a transferability attention mechanism. Experiments conducted on the UCF-HMDB and Kinetics-NEC Drone datasets, with different backbones, like I3D and STAM, show that TransferAttn outperforms state-of-the-art approaches. Also, we demonstrate that our DTAB yields performance gains when applied to other ViT-based methods for video UDA.**

I. Introduction

Among video analysis tasks, action recognition is both popular and challenging due to the many variations in how actions can be performed and captured, including differences in speed, duration, camera angles, actor movement, and occlusion [1].

Various deep learning methods for action recognition handle the temporal dimension differently. Some use 3D models to capture both spatial and temporal features, while others separate spatial and temporal data or use Recurrent Neural Networks (RNNs) to model temporal dynamics [2]. High human costs arise from the need for extensive video annotations to achieve good results, which is difficult and labor-intensive across many application domains [3].

Unsupervised Domain Adaptation (UDA) helps reduce the cost of manual data annotation by training models on labeled data from a source domain and unlabeled data from a target domain. Approaches like adversarial-based and transformer-based methods, have shown promising results [4]. However, only a few recent studies [1], [5]–[7] address video UDA for action recognition due to its more challenging aspects. Even fewer explore Vision Transformer (ViT) architectures [1].

Although previous work has pushed the performance of video UDA, there is still much room for improvement. Unlike most approaches that align frames with larger spatial and temporal domain gaps, we hypothesize that frames with smaller domain gap have a more meaningful action representation and can improve the adaptability.

Motivated by this assumption, we propose a novel method for video UDA, called Transferable-guided Attention (TransferAttn), which improves the potential of ViT. Our method uses a transformer encoder 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.

The main contributions of this paper are: (*i*) we propose DTAB, a novel transferable transformer for UDA, which uses a new attention mechanism capable of improving adaptation and domain transferability in ViT; and (*ii*) we conduct experiments on the UCF $\leftrightarrow$ HDMB$_{full}$ [8] and Kinetics $\rightarrow$ NEC-Drone [9] bechmarks, setting a new state-of-the-art result.

II. Related Work

Action Recognition: Deep learning and large-scale video datasets, like Kinetics, Moments-In-Time, and YouTube 8M, have greatly advanced action recognition. Deep learning-based solutions for action recognition are categorized by how how they model temporal dynamics into: (*i*) space-time networks, like C3D and I3D, which use a 3D convolutions; (*ii*) multi-stream networks, like TSN and TDN, which use optical flow data separately; and, (*iii*) hybrid models, like LSTMs [2].

Unsupervised Domain Adaptation: Various strategies for unsupervised domain adaptation (UDA) have been employed to alleviate domain shifts in images [10]. Adversarial-based methods are a standard choice and use a domain discriminator to minimize the domain gap through a min-max optimization, similar to Generative Adversarial Networks (GANs). In a different direction, metric-based methods aim to reduce the domain gap by learning domain-invariant features through discrepancy metrics, like Maximum Mean Discrepancy (MMD) and Joint Adaptation Networks (JAN) [10]. Driven by the success of ViTs, TVT [11] employs a transferability metric as a weight in the class token.

Unsupervised Domain Adaptation for Action Recognition: Despite the potential applications of UDA for action recognition, only a few recent studies have addressed this challenge [4]. For instance, MA^{2}LT-D [7] generates multi-level temporal features with domain discriminators. CleanAdapt [6] addresses source-free video domain adaptation by using noisy labels from a pre-trained source model. TranSVAE [5] alleviates spatial and temporal domain shift with latent factor constraints. Although transformers have achieved state-of-the-art performance in many tasks, for video UDA, they have been explored only in a few works, like UDAVT [1], which uses STAM [12] as backbone and performs domain alignment with a loss based on the Information Bottleneck (IB) principle [13].

III. Our Approach

In Section III-A, we describe our baseline model, which relies on a ViT architecture and adversarial domain adaptation. Then, in Section III-B, we introduce our domain transferable-guided attention block, called DTAB, and its components.

A. Baseline Model

Given the success of ViTs in diverse computer vision tasks [14], we built a baseline architecture that leverages a ViT encoder to produce better spatio-temporal features and an adversarial-based method for UDA, as shown in Fig. 1.

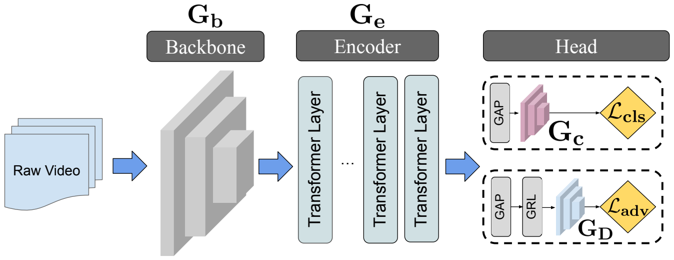

Fig. 1. A baseline architecture trained with an adversarial loss $\mathcal{L}_{adv}$ and a classification loss $\mathcal{L}_{cls}$. The backbone G_b extracts frame-level features and the encoder G_e learns meaningful semantic spatio-temporal representations.

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) consists of L layers, h heads and a hidden size of d. A patch embedding (G_p) may be used before the encoder to map features from the backbone to the encoder's input size.

Let $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ be a labeled source domain sharing the same set $\mathcal{Y}$ of categories with an unlabeled target domain $\mathcal{T} = \{x_i^t\}_{i=1}^{N_t}$. We denote the j-th frame from the i-th video as $x_{i,j}^s$ for the source domain and $x_{i,j}^t$ 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 $F_{i,j}^s$, and $F_{i,j}^t$ for the target domain. Also, we denote f_i^s and f_i^t 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 MLP classifier (G_C), that is not trained. 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 through the classification ($\mathcal{L}_{cls}$) and entropy ($\mathcal{L}_H$) losses, as given by:

$$\mathcal{L}_{cls} = -\frac{1}{N_s}\sum_{i=1}^{N_s} y_i^s \cdot \log G_C(f_i^s) \tag{1}$$

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

The adaptation branch is composed of a Gradient Reversal Layer (GRL) followed by a domain discriminator (G_D), which is a MLP classifier to identify the video domain. With the inversion of the gradients in the GRL, the encoder G_e learns to deceive the domain discriminator G_D, playing a min-max game, using on a adversarial loss ($\mathcal{L}_{adv}$), defined as:

$$\mathcal{L}_{adv} = -\frac{1}{N_s}\sum_{i=1}^{N_s} \log G_D(\mathrm{GRL}(f_i^s)) - \frac{1}{N_t}\sum_{i=1}^{N_t} \log G_D(\mathrm{GRL}(f_i^t)) \tag{3}$$

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

B. TransferAttn: Transferable-guided Attention

The baseline architecture with standard ViT encoder, as shown in Tables I and II, achieves limited improvement. We hypothesize that the self-attention does not fully leverage the transferable capabilities of ViT, as it it does not consider inter-domain specificity and focus only on intra-domain similarity.

To support our hypothesis, we propose TransferAttn, an improvement of our baseline model. The key difference between TransferAttn and our baseline model is 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 Fig. 2(a).

Instead of using self-attention, which gives attention to similar tokens or patches (in our case, frames) only within a same domain, we propose the Multi-head Domain Transferable-guided Attention (MDTA). This mechanism weights frames on how difficult it is to predict which domain they belong to (see Fig. 2(b)). Unlike methods that prioritize frames with larger domain gaps, like MA2LT-D [7], the reasoning behind MDTA is to highlight portions of a video that are similar across domains. As an example, Fig. 2(c) shows how MDTA weights frames. 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 [13] to find a shared representation space between the source and target domains.

Multi-head Domain Transferable-guided Attention. 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 that are more similar across domains and still semantic meaningful.

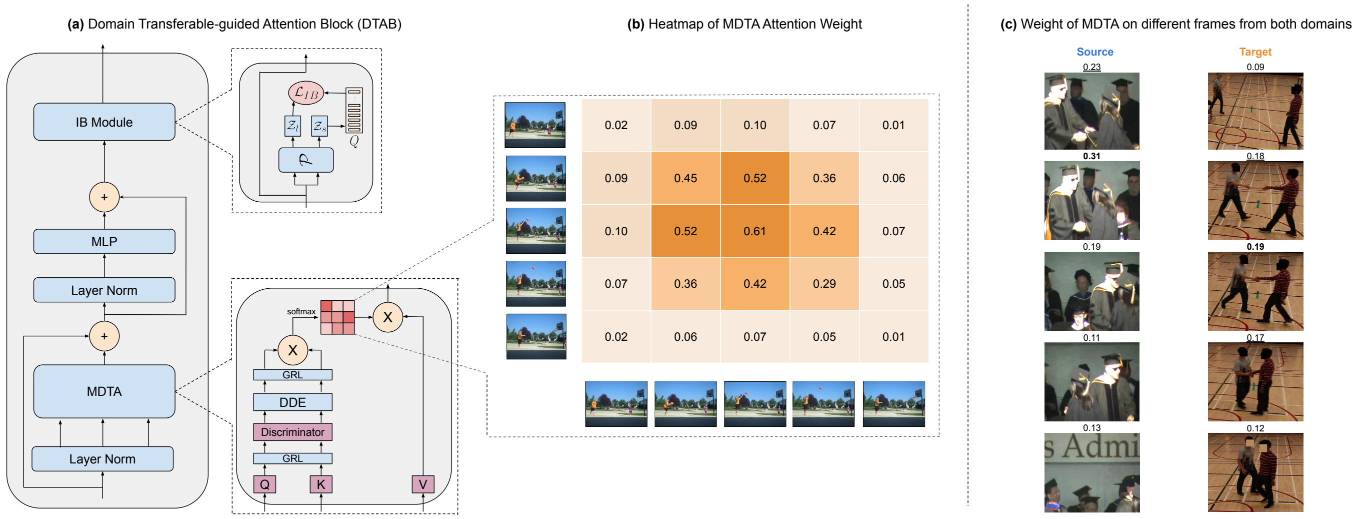

Fig. 2. DTAB overview. (a) The Domain Transferable-guided Attention Block follows a standard transformer block layout, except for our novel MDTA attention mechanism and the layer-wise Information Bottleneck (IB) calculation. (b) The heatmap of the transferable-attention weights, showing how MDTA focus on frames that has information that is more transferable between different domains and also brings more meaningful information about the ocurring action. (c) Temporal attention visualization compared between domains.

Similar to self-attention, MDTA is computed as in Eq. 4 and 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.

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

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

However, unlike self-attention, in MDTA, the attention scores are computed by taking the dot product between the vectors $Q^{'} = \text{GRL}(\text{DDE}(G_{D'}(\text{GRL}(QW_i^Q))))$ and $K^{'} = \text{GRL}(\text{DDE}(G_{D'}(\text{GRL}(KW_i^K))))$, where the Domain Discriminator Error (DDE) is a weighting function given by Eq. 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 a domain-separable feature space that is class-invariant.

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

This operation results in a matrix that defines which frames are more or less transferable, as depicted in Fig. 2(b). In other words, if DDE goes to one, it is more likely to deceive the domain discriminator $G_{D'}$ and should be more important when classifying the video action.

Information Bottleneck Module. To perform domain alignment, this module calculates a loss given by Eq. 7, where C is a cross-correlation matrix computed between mean centered representations from the source and target batches.

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

The loss 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 ($\mathcal{L}_{ib}$).

IV. Experiments and Results

We evaluated our approach through experiments on video UDA datasets and compared it to state-of-the-art methods.

A. Datasets

Experiments were conducted on 2 different benchmarks: **UCF** ↔ **HDMB**$_{full}$ [8] is one of the most widely used, containing a subset of videos from two datasets, UCF101 [15] and HMDB51 [16], representing 3,209 videos and 12 classes. **Kinetics** → **NEC-Drone** [9] contains videos from the Kinetics-600 [17] and NEC-Drone datasets, including a total of 10,118 videos and 7 classes.

B. Experimental Setup

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.

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

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

1) Results on UCF101↔HMDB51$_{full}$: As shown in Table I, we compare our results on different backbones. The first one is the I3D backbone, in which our approach surpasses even works that use multi-modal data (i.e, color and motion). From single-modal data, our approach achieves a significant average increase of 3.5% and, compared to multi-modal methods, can be seen as an average increase of 0.4%. For STAM backbone, our approach yields a average increase of 0.9%.

TABLE I
Classification Accuracy on UCF101$\leftrightarrow$ HMDB51$_{\text{FULL}}$.
Multi-modal methods are represented with (C + M).

Method	Backbone	U → H	H → U	Average
Source Only	I3D	80.6	89.3	85.0
MA^2LT-D [7]		89.3	91.2	90.3
TranSVAE [5]		87.8	99.0	93.4
CleanAdapt (C + M) [6]		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 [5]		93.5	99.5	96.5
UDAVT [1]		92.3	96.8	94.6
MA^2LT-D [7]		95.3	99.4	97.4
Baseline		93.4	98.9	96.1
TransferAttn (ours)		**97.2**	**99.7**	**98.5**

2) Results on Kinetics→NEC-Drone: Our approach was also evaluated in the Kinetics → NEC-Drone benchmark, as shown in Table II, achieving a significant increase of 9.5% in comparison with UDAVT, setting a new state-of-the-art result.

TABLE II
Classification Accuracy on Kinetics → NEC-Drone dataset.

Method	Backbone	K → N
Source Only	STAM	29.4
MA^2LT-D [7]		55.4
TranSVAE [5]		55.9
UDAVT [1]		65.3
Baseline		45.5
TransferAttn (ours)		**74.8**

D. Ablation Study

We studied the effect of our transformer block, DTAB, on other ViT-based methods for video UDA. We ran UDAVT [1] and reported the results of our reproduction using the author's code[1]. We also reported the results for the Transferability Adaption Module (TAM), which is the attention mechanism introduced in TVT [11]. As shown in Table III, adding our DTAB increased the accuracy by 1.7%, while TAM achieved marginal gains of less than 0.5%. These findings indicate that DTAB outperforms TAM in handling spatio-temporal data and can enhance other ViT-based methods for video UDA.

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

Method	U → H	H → U	Average
UDAVT (Our impl.) [1]	92.2	96.5	94.4
UDAVT + TAM [11]	92.5	96.9	94.7
UDAVT + DTAB	**94.2**	**97.9**	**96.1**

V. Conclusions

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 that were compared to ours, showing the effectiveness of our TransferAttn model. Our DTAB module also demonstrated to be a promising strategy when added to other transformer-based UDA methods, increasing their performance. As future work, we plan to augment our approach to utilize multi-modal data and make it able to integrate into source-free methods.

Acknowledgment

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

References

[1] 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 *ICPR*, 2022, pp. 1258–1265.

[2] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1366–1401, 2022.

[3] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, 2017.

[4] Y. Xu, H. Cao, Z. Chen, X. Li, L. Xie, and J. Yang, "Video unsupervised domain adaptation with deep learning: A comprehensive survey," *CoRR*, vol. abs/2211.10412, 2022.

[5] P. Wei, L. Kong, X. Qu, Y. Ren, zhiqiang xu, J. Jiang, and X. Yin, "Unsupervised video domain adaptation for action recognition: A disentanglement perspective," in *NeurIPS*, 2023.

[6] A. Dasgupta, C. V. Jawahar, and K. Alahari, "Overcoming label noise for source-free unsupervised video domain adaptation," *CoRR*, vol. abs/2311.18572, 2023.

[7] P. Chen, Y. Gao, and A. J. Ma, "Multi-level attentive adversarial learning with temporal dilation for unsupervised video domain adaptation," in *WACV*, 2022, pp. 776–785.

[8] M.-H. Chen, Z. Kira, G. Alregib, J. Yoo, R. Chen, and J. Zheng, "Temporal attentive alignment for large-scale video domain adaptation," in *ICCV*, 2019, pp. 6320–6329.

[9] J. Choi, G. Sharma, M. Chandraker, and J.-B. Huang, "Unsupervised and semi-supervised domain adaptation for action recognition from drones," in *WACV*, 2020, pp. 1706–1715.

[10] J. Li, L. Zhu, and Z. Du, *Unsupervised Domain Adaptation - Recent Advances and Future Perspectives*. Springer, 2024.

[11] J. Yang, J. Liu, N. Xu, and J. Huang, "Tvt: Transferable vision transformer for unsupervised domain adaptation," in *WACV*, 2023, pp. 520–530.

[12] G. Sharir, A. Noy, and L. Zelnik-Manor, "An image is worth 16x16 words, what is a video worth?" *CoRR*, vol. abs/2103.13915, 2021.

[13] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *IEEE Information Theory Workshop (ITW'15)*, 2015, pp. 1–5.

[14] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12 922–12 943, 2023.

[15] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[16] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *ICCV*, 2011, pp. 2556–2563.

[17] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *CoRR*, vol. abs/1808.01340, 2018.

[1] https://github.com/vturrisi/UDAVT (As of July, 2024)