Supplementary Materials for paper 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

A. Experimental Details

The adversarial training relies on the hyperparameter λ to control the strength of the Gradient Reversal Layer (GRL) on the adaptation head. Our DTAB module relies on the hyperparameters $\mathcal{Q}$ and α, where the first controls the queue size and the other controls the weight of the IB loss [3] and, in the GRL from DTAB, we fixed the weight λ_{DTAB} as 1 for every adaptation task. Also, we must define the batch size and the k sampling frames related to the training schedule.

For the **UCF** → **HMDB** [1] benchmark, we used a batch size of 32 and a sample of $k = 53$ frames. Due to its smaller size, we used a queue size $\mathcal{Q}$ of 512. Also, we used the IB loss of $\alpha = 0.001$ and adversarial loss of $\lambda = 1$. In the **HMDB** → **UCF** benchmark, the only change is in the adversarial loss of $\lambda = 0.5$ to make the training more stable.

For the **Kinetics** → **Gameplay** [1] benchmark, we used a batch size of 64 and a sample of $k = 23$ frames. In this adaptation task, we reduced the queue size $\mathcal{Q}$ to 512, used an IB loss of $\alpha = 0.001$, and a minor adversarial loss of $\lambda = 0.05$, making the training more stable.

In the **Kinetics** → **NEC-Drone** [2] benchmark, we used a batch size of 64, a sample of $k = 53$ frames, a queue size $\mathcal{Q}$ of 512, an IB loss of $\alpha = 0.025$, and an adversarial loss of $\lambda = 0.5$.

B. More Ablation Studies

This section reports the extra ablation studies conducted with our TransferAttn framework.

B.1. Effect of the DTAB position

To study the impact of the position of the DTAB module, we experimented by first changing all transformer blocks to DTAB, then changing only the first and last ones, and finally placing them in odd and even positions. The results in Table 7 show that our DTAB works better when used in the place of the last transformer block, where the patch features are more fine-grained than the others.

Table 7. Ablation study on Kinetics → NEC-Drone integrating the DTAB in different encoder positions.

DTAB Position	Backbone	K → N
All Blocks	STAM	36.0
First Only		38.2
Even Positions		54.0
Odd Positions		65.4
Last Only		**74.8**

B.2. Effect of the Fixed Classifier

One of the hypotheses we introduce in this paper is the use of a classifier with fixed random weights. This approach is motivated by the idea that fixing the classification boundaries forces the encoder G_e to learn a feature space that is more generalizable across domains, avoiding the classification head G_C to overfit on the source domain data. To study the impact of the fixed random classifier, we propose an ablation study to evaluate both our baseline and TransferAttn models with both learnable and fixed classifiers.

In Table 8, we present the accuracy results on the Kinetics → NEC-Drone benchmark, using the STAM backbone. As we can see, in both models, the use of a fixed random classifier yields an improvement in the final result, demonstrating that fixing the classification boundaries makes the encoder G_e to learn more robust features for UDA.

Table 8. Ablation study on Kinetics → NEC-Drone using learned and fixed classifier.

Method	Classifier	K → N
Baseline	Learnable	41.7
	Fixed	**45.5**
TransferAttn	Learnable	69.2
	Fixed	**74.8**

Table 9. Ablation study on Kinetics → NEC-Drone comparing how each class was impacted (class-wise accuracy).

Method	Walking	Running	Jumping	Drinking	Throwing an Obj.	Shaking Hands	Hugging
Baseline	0.0	100.0	0.0	21.7	33.3	77.8	96.2
+IB	29.7	80.7	0.0	73.9	75.0	74.1	96.1
+MDTA	89.1	3.9	42.3	100.0	72.2	70.4	100.0

B.3. Impacts of Adaptation on each Action Class

To study the impact of our MDTA mechanism, we conducted an experiment to analyze how the addition of our new attention mechanism impacts each type of class. As can be seen in Table 9, the baseline method achieved good results in a small set of classes and performed poorly on the rest. While the addition of IB improved recognition for certain classes, the best results were obtained when MDTA was combined with IB and the baseline method (last row).

From another perspective, we grouped the action classes by type, such as Pose (e.g., Walking, Running, and Jumping), Person-Object Interaction (e.g., Drinking and Throwing an Object), and Person-Person Interaction (e.g., Shaking Hands and Hugging), as shown in Table 10. Notably, the groups most impacted were Person-Object Interaction and Pose. Our main hypothesis is that MDTA's ability to focus on semantically meaningful frames enables the model to concentrate on frames containing more relevant information for classification, such as object movements or changes in pose.

Table 10. Ablation study on Kinetics → NEC-Drone comparing how each type of action was impacted (class-wise accuracy).

Method	Pose	Person-Object	Person-Person
Baseline	33.3	27.5	87.0
+IB	36.8	74.4	85.1
+MDTA	45.1	86.1	85.2

B.4. Domain Gap Limitations

The limitations of our approach were explored using the Kinetics→Gameplay dataset, which represents an adaptation from virtual data to real-world data. As shown in Table 2, our approach improves the SOTA results, even when applied on a combination of real and synthetic data.

Although we explored adaptation from synthetic to real data, adaptation to first-person or egocentric videos was not addressed. For future work, we plan to investigate the use of TransferAttn on the Jester [5] and Epic-Kitchens [4] datasets.

References

[1] 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] 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. 1

[3] 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

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In European Conference on Computer Vision (ECCV), 2018. 2

[5] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pages 2874–2882, 2019. 2