Balancing Known and Unknown in Open-Set Domain Adaptation via Attention

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

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

***Abstract*—Open-set unsupervised domain adaptation (OS-UDA) for video action recognition is a critical yet largely underexplored problem. Real-world applications must be robust to distribution shifts between training and deployment data (domain gap) and capable of identifying actions not seen during training (open-set). We introduce OS-DTAB, a lightweight, plug-and-play Vision-Transformer block designed to enhance existing OS-UDA frameworks. OS-DTAB replaces the standard clip-aggregation mechanism with a domain-transferable attention module. This component compels the model to focus on spatio-temporal cues that are simultaneously transferable and open-set aware. Our experiments on the HMDB↔UCF benchmark show that OS-DTAB sets a new state-of-the-art, achieving a Harmonic Open-Set (HOS) score that surpasses models with more robust backbones.**

I. Introduction

Action recognition is a cornerstone problem in video analysis, yet it remains challenging due to intra-class variations in speed, viewpoint, and background clutter. While deep learning models have achieved remarkable success, their reliance on large-scale labeled datasets limits their scalability. Unsupervised Domain Adaptation (UDA) mitigates this by transferring knowledge from a labeled source domain to an unlabeled target domain. However, UDA methods operate under the closed-set assumption, where both domains share the same set of classes.

In realistic scenarios, a target environment will inevitably contain new, previously unseen actions. This gives rise to the Open-Set UDA (OS-UDA) setting, where a model must not only classify instances from shared classes correctly but also identify and reject instances from unknown classes. This task is significantly more challenging as it requires the model to balance domain invariance for shared classes with sensitivity to novel patterns for unknown ones, a process often hampered by negative transfer from the unknown classes. Despite its practical importance, OS-UDA for video remains largely underexplored, with only a few methods like COLOSEO [1], AutoLabel [2], and CEVT [3] addressing it directly.

Motivated by these challenges, we introduce the Open-Set Domain-Transferable-guided Attention Block (OS-DTAB). OS-DTAB is a lightweight, plug-and-play module designed to enhance existing video OS-UDA frameworks. In our work, we integrate it into COLOSEO [1], replacing its simple MLP-based clip aggregator. Our main contributions are threefold: (i) we propose OS-DTAB, an extension of the Domain-Transferable Attention Block (DTAB) [4], specifically tailored for the OS-UDA setting; (ii) we demonstrate that OS-DTAB acts as an effective plug-and-play component that improves the performance of a baseline model; and (iii) we establish a new state-of-the-art result on the HMDB↔UCF$_{\text{full}}$ [5] benchmark.

II. Related Work

Action Recognition: Deep-learning approaches have achieved remarkable performance in human action recognition from videos. These methods can be categorized into: (i) *space-time networks*, like C3D and I3D, employ 3D convolutions to learn joint spatial and temporal representations; (ii) *Multi-stream networks*, such as TSN and TDN, process RGB frames and optical flow in separate branches and fuse their predictions to leverage appearance–motion cues; (iii) *hybrid* CNN–RNN architectures incorporate recurrent units on top of convolutional features to capture longer temporal dependencies.

Open-set Unsupervised Domain Adaptation: Open-set UDA extends the conventional UDA paradigm by incorporating unknown classes in the target domain. The objective is to train a model that correctly classifies the shared classes while rejecting samples from any other class into a single "unknown" category. Early methods focused on adversarial strategies, like OSBP [6], or thresholding methods. More recent approaches, such as OVANet [7] and its extensions [8], iteratively refine decision boundaries; while others, like HyMOS [9], employ clustering strategies to improve feature space compactness.

Open-set Unsupervised Video Domain Adaptation: The video counterpart of OS-UDA remains largely underexplored. COLOSEO [1] introduces a temporal contrastive mechanism to simultaneously separate unknown classes while tightening clusters of known ones. More recently, AutoLabel [2] detects scene objects using ViLT and leverages CLIP [10] to differentiate open-set from closed-set classes, forming object clusters that guide subsequent adaptation.

III. Background

In this section, we outline the foundational methods upon which our approach is built.

A. COLOSEO Model

COLOSEO [1] is a recent state-of-the-art framework for open-set unsupervised video domain adaptation that simplifies

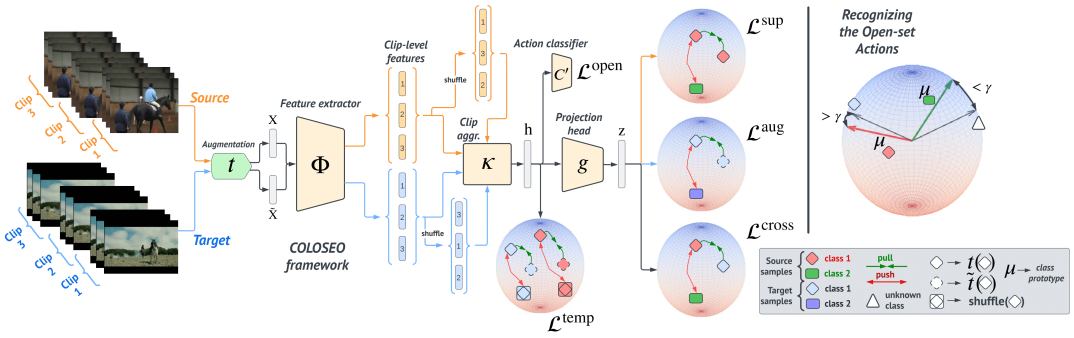

Fig. 1. COLOSEO architecture. The default architecture uses MLP to fuse multi-clip features and contrastive learning to learn better shared and private feature representation (reproduced from [1])

the adaptation process through a unified contrastive learning strategy. Figure 1 illustrates the architecture of the method.

Unlike adversarial alignment approaches, COLOSEO [1] learns compact and well-clustered video feature representations that implicitly align shared action classes across source and target domains while segregating unknown (target-private) classes. Furthermore, COLOSEO introduces a video-oriented temporal contrastive loss ($\mathcal{L}^{\text{temp}}$) by contrasting correct versus shuffled clip orders, thereby leveraging temporal dynamics to improve robustness to unknown classes.

In its design, COLOSEO divides each video into a fixed number of short clips, which are processed by a 3D CNN backbone (Φ) to extract clip-level features. These features are then aggregated into a single video-level representation by a lightweight network ($\mathcal{K}$). Specifically, the original COLOSEO implementation employs a simple two-layer MLP as the clip aggregator module: this MLP fuses multiple clip feature vectors into a single 1024-dimensional video embedding. While effective for pooling information, this MLP-based aggregation does not explicitly model temporal relationships among clips.

In our work, we build upon the COLOSEO [1] framework by replacing the MLP-based aggregator with a novel attention-based mechanism. This modification allows the model to dynamically weight and integrate information from each clip, thereby capturing richer temporal cues and further improving adaptation performance.

B. DTAB Module

The Domain Transferable-guided Attention Block (DTAB) [4] was introduced as a key component of the TransferAttn framework [4] for video unsupervised domain adaptation. DTAB modifies the standard self-attention mechanism into a transferable attention mechanism, encouraging a Vision Transformer (ViT) to focus on features shared across domains. Concretely, each token (patch-level feature) is evaluated by a discriminator to predict whether it originates from the source or target domain. Tokens that confuse the discriminator (i.e., are difficult to classify as source or target, indicating domain-invariance) are assigned higher attention weights, prompting the model to prioritize domain-generalizable patterns.

Beyond the transferable attention mechanism, DTAB integrates an information bottleneck (IB) principle to further align latent features. The IB component introduces a regularization loss that minimizes discrepancies between source and target feature distributions, defined in Equation 1, where C is the cross-correlation matrix between the projections of source and target samples.

$$\mathcal{L}_{\text{IB}} = \sum_{i=1}^{m}(1 - C_{i,i})^2 + \lambda \sum_{i=1}^{m}\sum_{j \neq i}^{m}(C_{i,j})^2 \tag{1}$$

The original DTAB [4] was designed for closed-set unsupervised domain adaptation, and we observed empirically that it underperforms when applied to open-set tasks.

To address this limitation, we introduce a new loss component specifically tailored to OS-UDA settings. Our goal is to preserve DTAB's ability to learn transferable features while equipping it to detect and manage unknown-class data in the target domain. By adapting DTAB to the open-set scenario, we aim to retain its effective domain-invariant attention weighting and feature alignment, thereby bridging the gap between closed-set success and open-set applicability.

This forms the foundation of our proposed method in the following sections.

IV. Our Approach: Open-set DTAB

In this work, we extend the Domain-Transferable-guided Attention Block (DTAB) [4] to the open-set unsupervised domain adaptation (OS-UDA) setting, and apply this new method as a clip aggregator on COLOSEO [1]. The resulting architecture, termed Open-set Domain-Transferable-guided Attention Block (OS-DTAB), shown in Figure 2, introduces two complementary mechanisms: (i) **Shared-class alignment**, encouraging features of known classes to lie in a domain-invariant sub-space; (ii) **Private-class separation**, explicitly pushing unknown target classes away from the source manifold to mitigate negative transfer. In the following, we detail the key components that comprises this idea:

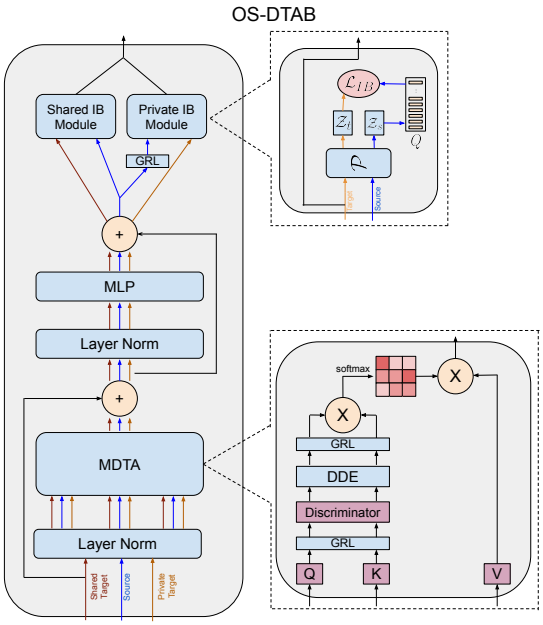

Fig. 2. Architecture of OS-DTAB. OS-DTAB follows a standard transformer architecture, except for the new attention mechanism and our new open-set IB module (adapted from [4])

Target partition via open-set classifier: Following COLOSEO [1], we attach an auxiliary head C' that is explicitly trained to label target-private instances as `unknown`. The scores produced by C' enable us to partition every target mini-batch into a **Shared** subset samples that are likely to belong to classes seen in the source and a **Private** subset comprising the remaining, unseen classes. This split will be used in the following subsections

Shared Information Bottleneck: Given source features (z_s) and shared-target features (z_t^{sh}), we adapt the IB loss originally used in DTAB, described in Equation 2, where C is the cross-correlation matrix computed over a batch and an accompanying queue that stores additional source features. Minimizing $\mathcal{L}_{\text{IB}}^{\text{sh}}$ maximizes agreement between domains only for classes deemed shared, yielding class-conditional, domain-invariant representations.

$$\mathcal{L}_{\text{IB}}^{\text{sh}} = \sum_{i=1}^{m}(1-C_{i,i})^2 + \lambda \sum_{i=1}^{m}\sum_{j\neq i}^{m}(C_{i,j})^2, \quad C = \text{corr}(\boldsymbol{z}_s, \boldsymbol{z}_t^{\text{sh}}), \tag{2}$$

Private Information Bottleneck: To actively diverge private target classes from the source, we feed source features through a Gradient Reversal Layer (GRL) before calculating the IB loss, as described in Equation 3.

$$\boldsymbol{z}_s^* = \text{GRL}(\boldsymbol{z}_s), \qquad \mathcal{L}_{\text{IB}}^{\text{pr}} = \mathcal{L}_{\text{IB}}\left(\boldsymbol{z}_s^*, \boldsymbol{z}_t^{\text{pr}}\right). \tag{3}$$

Because the GRL multiplies the source gradient by a negative factor, minimizing $\mathcal{L}_{\text{IB}}^{\text{pr}}$ maximizes the discrepancy between $\boldsymbol{z}_s$ and private-target features (z_t^{pr}), effectively pushing unknown classes away from the shared sub-space.

Overall Loss: Our full objective, shown in Equation 4, combines the standard COLOSEO losses with the two IB terms introduced above. The hyperparameter λ_{pr} and λ_{sh} control the trade-off between alignment and separation:

$$\mathcal{L} = \underbrace{\mathcal{L}^{\text{open}} + \mathcal{L}^{\text{sup}} + \mathcal{L}^{\text{aug}} + \mathcal{L}^{\text{cross}} + \mathcal{L}^{\text{temp}}}_{\text{COLOSEO Loss}} + \underbrace{\lambda_{\text{pr}}\mathcal{L}_{\text{IB}}^{\text{pr}} + \lambda_{\text{sh}}\mathcal{L}_{\text{IB}}^{\text{sh}}}_{\text{Our loss}} \tag{4}$$

V. Experiments and Results

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

A. Datasets

Experiments were conducted on the UCF↔HMDB$_{full}$ benchmark [5], one of the most widely-used datasets for video domain adaptation. For the OS-UDA setting, this dataset is adapted to contain 6 shared classes and 6 private classes, resulting in a total of 3,209 videos.

B. Evaluation Protocol

To evaluate our proposed method to the baselines, we adopt 3 different metrics proposed by CEVT [3]. The first one is **OS*** known as the accuracy over the shared classes. Also, **UNK** is reported as the correct ratio of private classes predicted as "unknown". And finally, the **HOS**, described in the Equation 5, which is the preferred metric in open-set literature, due to its emphasis on balancing **OS*** and **UNK**.

$$HOS = 2 \cdot \frac{OS^* \cdot UNK}{OS^* + UNK} \tag{5}$$

C. Implementation details

The backbone of COLOSEO is I3D, pretrained on Kinetics-400. We extract 3 clip-level features uniformly sampled from each video. Following COLOSEO [1], we use the SGD optimizer with a learning rate of 0.01 and a batch size of 16. The hyperparameters introduced in this work for $\mathcal{L}_{\text{IB}}^{\text{sh}}$ and $\mathcal{L}_{\text{IB}}^{\text{pr}}$ are set to 0.001 and 0.1, respectively.

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

As shown in Table I, our method achieves the highest **HOS** score on both adaptation tasks, establishing a new state-of-the-art. Indeed, it demonstrates a much better balance between the shared (**OS***) and private (**UNK**) class accuracies. Notably, our method surpasses AutoLabel [2], a method that leverages the powerful CLIP backbone pre-trained on extensive language-vision data. Our I3D-based model, which does not use such external supervision, outperforms it by a large margin.

TABLE I
Classification results on State-of-the-art Methods

Method	Backbone	HMDB → UCF			UCF → HMDB		
		OS*	**UNK**	**HOS**	**OS***	**UNK**	**HOS**
OVANet [7]	I3D	90.4	76.0	82.6	67.0	71.4	69.6
CEVT [3]	I3D	59.7	93.0	72.7	60.3	96.1	74.1
COLOSEO [1]	I3D	81.1	88.7	84.7	76.7	**98.9**	86.4
AutoLabel [2]	CLIP	82.5	**94.3**	88.0	82.9	88.2	85.5
Ours	I3D	**94.3**	92.1	**93.2**	**84.7**	97.8	**90.8**

E. Ablation Study

We conduct ablation studies to validate our design choices. **Impact of Clip Aggregator:** In Table II, we compare our proposed OS-DTAB with other clip aggregation strategies. The baseline MLP from COLOSEO performs reasonably well. Replacing it with a standard Transformer encoder improves temporal modeling, boosting **OS*** but hurting **UNK** slightly. Using the original closed-set DTAB drastically improves **OS*** to 94.0% but causes a significant drop in **UNK** accuracy (from 88.7% to 81.2%), as it mistakenly forces alignment for all classes. Only our proposed OS-DTAB, with its dual-loss mechanism, is able to improve both **OS*** and **UNK**, leading to the best overall **HOS** score.

TABLE II
RESULTS COMPARING DIFFERENT CLIP AGGREGATION TECHNIQUES

Clip Aggregator	HMDB → UCF			UCF → HMDB		
	OS*	**UNK**	**HOS**	**OS***	**UNK**	**HOS**
MLP	81.1	88.7	84.7	76.7	98.9	86.4
Transformer	88.3	87.3	87.8	78.6	98.9	87.6
DTAB	94.0	81.2	87.1	84.3	87.1	85.7
OS-DTAB	**94.3**	**92.1**	**93.2**	**84.7**	**97.8**	**90.8**

Impact of Loss Components: In Table III, we analyze the effect of our two proposed IB losses. Using only the shared loss ($\mathcal{L}_{\text{IB}}^{\text{sh}}$) improves shared-class accuracy (**OS***) at the expense of unknown-class detection (**UNK**), confirming its role in domain alignment. Conversely, using only the private loss ($\mathcal{L}_{\text{IB}}^{\text{pr}}$) improves **UNK** but degrades **OS***, confirming its role in separating the feature spaces. A synergistic effect is observed only when both losses are combined, leading to a substantial +6.4 point improvement in **HOS** over the baseline without these losses. This demonstrates that both components are necessary to reconcile the trade-off between known-class alignment and unknown-class separation.

TABLE III
RESULTS COMPARING THE IMPACT OF EACH LOSS

#	$\mathcal{L}_{\text{IB}}$	$\mathcal{L}_{\text{IB}}^{sh}$	$\mathcal{L}_{\text{IB}}^{pr}$	HMDB → UCF		
				OS*	**UNK**	**HOS**
(a)				86.3	87.3	86.8
(b)	✓			94.0 (+7.7)	81.2 (-6.1)	87.1 (+0.3)
(c)		✓		95.2 (+8.9)	79.4 (-7.9)	86.6 (-0.2)
(d)			✓	86.4 (+0.1)	89.7 (+2.4)	88.0 (+1.2)
(e)		✓	✓	94.3 (+8.0)	92.1 (+4.8)	93.2 (+6.4)

VI. CONCLUSIONS

In this work, we addressed the largely unexplored problem of OS-UDA by enhancing the DTAB module and integrating it into the COLOSEO framework [1].

Our resulting architecture, OS-DTAB, introduces (i) an attention-based clip aggregator that captures richer temporal cues than the original MLP, and (ii) two complementary information-bottleneck losses that simultaneously align shared classes and repel target-private ones. Together, these components deliver a principled balance between domain invariance and unknown-class separation.

Experiments on the HMDB↔UCF benchmark [5] demonstrate the effectiveness of our contributions: OS-DTAB achieves a new state-of-the-art **HOS** of 93.2% on HMDB→UCF and 90.8% on UCF→HMDB, surpassing baselines such as COLOSEO [1] and AutoLabel [2]. Ablation studies confirm the necessity of both components: the attention-based aggregator enhances temporal reasoning, while the dual bottleneck losses reconcile the trade-off between known-class accuracy (**OS***) and unknown-class detection (**UNK**).

For future work, we plan to extend OS-DTAB to more challenging benchmarks and investigate its application to other vision tasks, such as image-based open-set domain adaptation.

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] G. Zara, V. d. C. Turrisi, S. Roy, P. Rota, and E. Ricci, "Simplifying open-set video domain adaptation with contrastive learning," *Computer Vision and Image Understanding*, vol. 241, p. 103953, 2024.

[2] G. Zara, S. Roy, P. Rota, and E. Ricci, "Autolabel: Clip-based framework for open-set video domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 11 504–11 513.

[3] Z. Chen, Y. Luo, and M. Baktashmotlagh, "Conditional extreme value theory for open set video domain adaptation," in *Proceedings of the 3rd ACM International Conference on Multimedia in Asia (MMAsia)*, 2022.

[4] A. Sacilotti, S. F. dos Santos, N. Sebe, and J. Almeida, "Transferable-guided attention is all you need for video domain adaptation," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 8680–8690.

[5] M.-H. Chen, Z. Kira, G. Alregib, J. Yoo, R. Chen, and J. Zheng, "Temporal attentive alignment for large-scale video domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6320–6329.

[6] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, "Open set domain adaptation by backpropagation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 606–621.

[7] K. Saito and K. Saenko, "Ovanet: One-vs-all network for universal domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9000–9009.

[8] L. F. A. e Silva, S. F. dos Santos, N. Sebe, and J. Almeida, "Beyond the known: Enhancing open set domain adaptation with unknown exploration," *Pattern Recognition Letters*, vol. 189, pp. 265–272, 2025.

[9] S. Bucci, F. C. Borlino, B. Caputo, and T. Tommasi, "Distance-based hyperspherical classification for multi-source open-set domain adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1119–1128.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.