LCMV: Lightweight Classification Module for Video Domain Adaptation

LCMV: Lightweight Classification Module for Video Domain Adaptation

Julian Neubert (✉)[1][0009−0002−7482−9789], Mirco Planamente[1,2,3][0000−0001−7238−1867], Chiara Plizzari[1][0000−0003−4984−7432], and Barbara Caputo[1,3][0000−0001−7169−0158]

[1] Politecnico di Torino
[2] Istituto Italiano di Tecnologia
[3] CINI Consortium
julian.neubert@gmx
{mirco.planamente,chiara.plizzari,barbara.caputo}@polito.it

Abstract. Video action recognition models exhibit high performance on in-distribution data but struggle with distribution shifts in test data. To mitigate this issue, Unsupervised Domain Adaptation (UDA) methods have been proposed, consisting in training on labeled data from a *source* domain and incorporating unlabeled test data from a *target* domain to reduce the domain gap. This requires simultaneous access to data from both domains, which may not be practical in real-world scenarios due to privacy issues. A more practical approach called Source-Free Domain Adaptation (SFDA) has been recently proposed, which consists in adapting a well-trained source model using only unlabeled target data. However, existing SFDA methods are computationally intensive and designed for specific architectures. In this paper, we propose an approach called Lightweight Classification Module for Video Domain Adaptation (LCMV). LCMV is based on a backpropagation-free prototypical algorithm, which efficiently adapts a source model using unlabeled target data only. Results on two popular datasets, HMDB-UCF$_{full}$ and EPIC-Kitchens-55, show significant improvements of LCMV compared to the previous state-of-the-art SFDA methods, and competitive results when compared to state-of-the-art UDA methods.

Keywords: Source-Free Domain Adaptation · Action Recognition

1 Introduction

Video action recognition is attracting an ever-growing amount of interest in the research community [18,2]. However, despite the advances in the field, models still suffer from the so-called "environmental bias" [24]. In action recognition this problem is amplified by the high-dimensional data and complex environments of videos, causing a significant drop in classification accuracy when evaluating models in new domains (see Figure 1).

So far, most researchers address this issue in the Unsupervised Domain Adaptation (UDA) setting, using labeled samples from the *source* domain and unlabeled samples from the *target* domain to reduce the models' bias [3,18] and

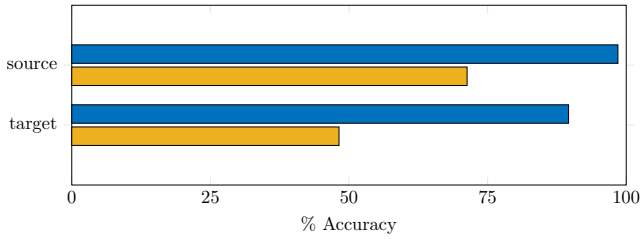

Fig. 1. Evaluating a model in a new environment often leads to significantly reduced performance, highlighting the need for domain adaptation methods. This Figure shows the prediction accuracy on **UCF-HMDB**$_{full}$ and **EPIC-Kitchens-55** when evaluating our I3D model on a seen source and unseen target domain without any adaptation.

achieve optimal performance on the target domain. These methods have proven to be successful in improving the robustness of models and enabling them to perform well across different environments. However, a key drawback of UDA is that it requires access to both source and target data during training, which may not be possible in many real-world applications where data sharing is restricted due to privacy concerns. In the Source-Free Domain Adaptation (SFDA) setting, which presents a more challenging yet realistic scenario, the training and adaptation processes are separated, allowing to adapt pretrained off-the-shelf source models using exclusively unlabeled target data. Current SFDA methods for action recognition achieve this by training with an ensemble of complex loss functions, for example, to ensure internal consistency [28] or align feature representations [11] and adapt the model using the weighted sum of many individual loss functions [4,11,28].

However, these SFDA approaches present several new challenges. Firstly, determining the individual weight for each component of the loss function typically entails manual selection of hyperparameters, necessitating adjustments on a per-dataset basis. Moreover, conducting backpropagation-based training during the adaptation phase proves computationally demanding, demanding large amounts of high-quality target data, and potentially compromising the performance of the underlying model. Lastly, numerous SFDA methods impose tight architectural constraints, tailored to specific architectures [28], thereby limiting modifications to the source model [4] or necessitating the design of unique architectures [11].

In this work, we propose a Lightweight Classification Module for Video Domain Adaptation (LCMV). Specifically, LCMV replaces the final classification layer with a prototypical-based classification module. Each class is represented by a "class prototype", and classification is done by assigning samples to the nearest prototype in feature space, without requiring any optimization process.

To summarize, our contributions are threefold:

- We develop a backpropagation-free architecture-independent classification solution for Source-Free Domain Adaptation for action recognition;
- Our method preserves the integrity of the original source model, allowing for seamless adaptation to multiple domains and adjustments to continuous domain shift;

– We validate our method on the UCF-HMDB$_{full}$ and EPIC-Kitchens-55 datasets, achieving state-of-the-art results.

2 Related Work

Action Recognition. The goal of Action Recognition (AR) is to classify the action that the subject is performing in a video. Architectures generally consist of convolutional networks utilizing either 2D [16,25,32] or 3D [1,18,20,27] convolutions. Sometimes additional modules are added to improve temporal modeling capabilities [9,16,19,32]. Finally, incorporating information from multiple modalities has proven to improve performance, especially in the cross-domain scenario [4,11,13,18,29].

Unsupervised Domain Adaptation for AR. Domain adversarial networks [10] have emerged as a well-established and widely used approach, integrated into numerous recent UDA methods for action recognition [26,3,6,18,29]. TransVAE [26] attempts to disentangle the domain information from task-specific information during the adaptation. TA^{3}N [3] integrates a temporal relation module for temporal alignment. Other methods use contrastive learning techniques [19,13]. Another research direction focuses on incorporating self-supervised learning as an auxiliary task to improve feature learning [6,18]. Finally, some works use pseudo-labels to guide the adaptation on unlabeled target data [19,13].

Source-Free Domain Adaptation. The application of SFDA methods to action recognition is a relatively new area of investigation. SFTADA [4] proposed to address SFDA in videos by learning temporal consistency to transfer robust centroid-based temporal attention weights from the source domain to the target domain. ATCoN [28] learns a temporal consistency which is composed of both feature and source prediction consistency. Recently, MTRAN [11] presented a loss derived from Mixup [31] to align the internal representation of source and target features, which is applied to their visual transformer-inspired architecture. CleanAdapt [8] proposed a self-training approach that selects the clean samples from the noisy pseudo-labeled target domain ones. To the best of our knowledge, no previous SFDA publication specific to action recognition exist, that is completely backpropagation-free or formulated independently of the underlying architecture.

Prototypical Classification. Prototypical networks have proven successful for zero-shot domain adaptation [21,5]. Subsequently, the use of prototypes has found its way into domain generalization, UDA, and SFDA, often to generate pseudo-labels [19,28,11,12] or to robustly estimate attention weights [4]. Inspired by [12], we propose to use prototypes directly as the model's prediction in the SFDA setting, where the larger amount of available data allows for significantly more robust prototypes. To the best of our knowledge, our work represents the

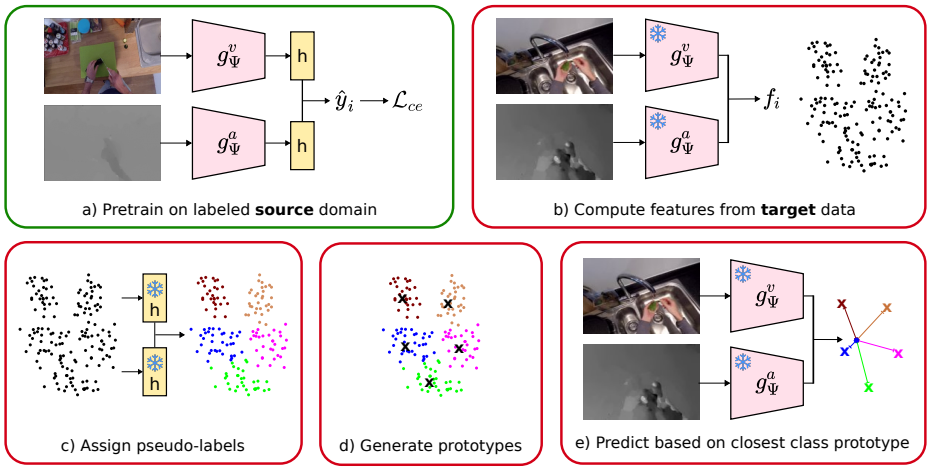

Fig. 2. Overview of LCMV. (a) Before adaptation, the source model (feature extractor and classification layer) is trained on the labeled source domain. (b) The source model is used to generate representations for target data from unlabeled target domain. (c) The source classifier is used to assign pseudo-labels for target. (d) Class prototypes are computed as mean feature vectors. (e) LCMV predicts samples based on the closest class prototype.

first attempt to apply this concept in SFDA for action recognition, and it introduces a backpropagation-free approach to the field.

3 Proposed Method

We propose to replace the standard classification layer of action recognition models with a backpropagation-free prototypical classifier which we call Lightweight Classification Module for Video Action Recognition (LCMV). Given a model pre-trained on source data, we feed it with samples from target and assign a class label to them. We then compute class prototypes as the normalized average feature vectors of target samples from each class. The final classification is performed by assigning each target sample to the closest class prototype. In this section, we formally describe the proposed LCMV and discuss it in detail.

3.1 Multi-Modal Source-Free Domain Adaptation

In the Unsupervised Domain Adaptation (UDA) setting, we are given a source domain $\mathcal{S}$, where $\mathcal{S} = \{(x_{s,i}, y_{s,i})\}_{i=1}^{N_s}$ is composed of N_s source samples with known labels $y_{s,i} \in Y_s$, and a target domain $\mathcal{T} = \{x_{t,i}\}_{i=1}^{N_t}$ of N_t target samples whose labels are unknown. The goal is to transfer knowledge from the labeled source domain to the unlabeled target domain by reducing the distribution discrepancy between the two during training. In the Source-Free Domain Adaptation (SFDA) setting we are only given a source model pre-trained on $\mathcal{S}$ with

supervision, but we do not have access to $\mathcal{S}$ for adaptation on target. The key idea is thus to learn discriminative target features while aligning them with the source data distribution embedded within the source classifier.

In our SFDA setting, video clips x_i from both domains consist of two modalities $x_i = \{x_i^v, x_i^f\}$, which are visual appearance (RGB) x_i^v and motion (optical flow) x_i^f respectively. Predictions are made by a two-stream action recognition model which averages the final softmaxed outputs over both modalities to return a fused probability score. For simplicity, we generically refer to both modalities in our equations, unless explicitly stated otherwise.

3.2 Prototype Generation

We are given a model trained on source data, where each modality stream $\Phi = h_{\{w,b\}} \circ g_\Psi$ can be divided into a feature extractor g_Ψ with parameters Ψ and a linear classification layer $h_{\{w,b\}}$ with weights w and bias b (see Figure 2-a). Prediction $\hat{y}_i$ for the input sample x_i to belong to class k is then given by

$$p(\hat{y}_i = k | x_i, \Psi, w, b) = \frac{\exp(d(w^k, g_\Psi(x_i)) + b^k)}{\sum_j \exp(d(w^j, g_\Psi(x_i)) + b^k)} \tag{1}$$

where w^k and b^k are the k-th row of the weight matrix and bias parameter of $h_{\{w,b\}}$ respectively, and $d(w^k, g_\Psi(x_i)) = w^k \cdot g_\Psi(x_i)$ is the dot product between w^k and $g_\Psi(x_i)$.

Let $\{x_1, \ldots, x_N\}$ be the set of all videos available in the training set of the target domain. Given an action video x_i we sample N_c clips $\{x_{i,1}, \ldots, x_{i,N_c}\}$ it. Using the given source model, we obtain the predictions over all clips $\{\hat{y}_{i,1}, \ldots, \hat{y}_{i,N_c}\}$ and average them to get a more robust prediction $\tilde{y}_i$ ("pseudo-label") for the video (Figure 2-c). Similarly, we use the normalized mean feature vector f_i over all clip features $\{g_\Psi(x_{i,1}), \ldots, g_\Psi(x_{i,N_c})\}$ to obtain a feature-based representation of video x_i (Figure 2-b).

Each prototype c^k for each class k is extracted by simply averaging the feature representation f_i of all videos x_i belonging to class $\tilde{y}_i = k$ (Figure 2-d):

$$c^k = \frac{1}{N_k} \sum_{\substack{i \\ \tilde{y}_i = k}} f_i \tag{2}$$

Following [12], we improve the prototypes quality by selecting for each class k the t videos with the lowest prediction entropy and compute the prototypes in Equation 2 only on this subset.

3.3 Final Prediction

To classify target samples using the computed prototypes, we simply assign them to the closest class prototype in feature space (Figure 2-e). Given a sample x_i we compute its class probabilities as

$$p(\hat{y}_i = k | x_i, \Psi) = \frac{\exp(d(c^k, g_\Psi(x_i)))}{\sum_j \exp(d(c^j, g_\Psi(x_i)))} \tag{3}$$

where $d(c^k, g_\Psi(x_i)) = (c^k - g_\Psi(x_i))$ is the Euclidean distance between c^k and $g_\Psi(x_i)$. Note that this formulation is the same as Equation 1, by substituting w_k with c^k and removing the bias term. Indeed, the rows w_k of the weight matrix of a standard classification layer can be seen as a class prototype [12].

Finally, we iteratively refine the prototypes similar to how the k-means algorithm works. This is accomplished by using the formulation described in Equation 3 to generate more accurate pseudo-labels $\tilde{y_i}$ for the target videos x_i. The prototypes are then updated accordingly using Equation 2.

The proposed LCMV offers a distinct advantage: our prototypes are not learned parameters that are susceptible to domain shift. Instead, they are directly computed from the target data, ensuring robustness and reducing the impact of the distribution gap.

4 Experiments

We evaluate the ability of LCMV to perform well on target data by comparing it against baseline and state-of-the-art Unsupervised Domain Adaptation (UDA) and Source-Free Domain Adaptation (SFDA) methods adapted to our setting. We then perform an analysis of the impact of class imbalance and multi-modal learning on our method. Finally, we show ablations on its different components.

4.1 Experimental Setting

Datasets. We evaluate our proposed method on the two most commonly used datasets for domain adaptation for action recognition: **EPIC-Kitchens-55** [7] and **UCF-HMDB**$_{full}$ [23,14]. **EPIC-Kitchens-55** contains egocentric action videos recorded in different kitchens across the world. We follow the setting proposed in [18], including the three largest kitchens by number of samples, each representing one domain (D1, D2, and D3). For **UCF-HMDB**$_{full}$ we follow the setting proposed in [3], which consists of 12 overlapping classes between the UCF101 and HMDB51 datasets. We train the model on one dataset and evaluate it on the other. In the following we use the abbreviation H→U to denote training on HMDB51 and evaluating on UCF101, and vice versa.

Implementation Details. Our model uses a two-stream I3D backbone pretrained on Kinetics [1] with averaged late fusion to combine multi-modal inputs. Flow frames are extracted using the TV-L1 algorithm [30]. Clips consist of 16 frames following the dense sampling strategy with random cropping, scale jitters, and horizontal flipping as augmentations. We evaluate our model using the average prediction on 5 clips per video. On UCF-HMDB$_{full}$ prototypes are computed using the 20 lowest entropy samples and refined through a single iteration of the k-means algorithm. On EPIC-Kitchens-55 we do not perform any prototype adjustment. Experiments have been run on a workstation with two NVIDIA GeForce GTX 1070 GPUs.

Table 1. Comparison with DA and SFDA methods on UCF-HMDB$_{full}$. Best results are marked in **bold** and MM indicates multi-modal results.

Method	Backbone	SFDA	MM	U→H	H→U	Avg
ATCoN [28]	ResNet-101	✓	-	79.7	85.3	82.5
SFTADA [4]	ResNet-101	✓	✓	87.2	91.2	89.2
STCDA [22]	I3D	-	-	83.1	92.1	87.6
CoMix [19]	I3D	-	-	86.7	93.9	90.3
CIA + TA3N [29]	I3D	-	-	91.9	94.6	93.2
TransVAE [26]	I3D	-	-	87.8	99.0	93.4
TA3N [3,6]	I3D	-	-	81.4	90.5	86.0
SAVA [6]	I3D	-	-	82.2	91.2	86.7
MM-SADA [18,13]	I3D	-	✓	84.2	91.1	87.7
CIA [29]	I3D	-	✓	90.6	94.2	92.4
Kim et al. [13]	I3D	-	✓	84.7	92.8	88.8
SHOT [15,11]	I3D	✓	✓	89.7	91.8	90.8
MTRAN [11]	I3D	✓	✓	**92.2**	95.3	93.8
CleanAdapt [8]	I3D	✓	✓	89.8	**99.2**	94.5
Ours	I3D	✓	✓	**92.2**	99.0	**95.6**

4.2 Comparison with the State-of-the-Art

We compare LCMV results with state-of-the-art works in both UDA and SFDA settings. While UDA methods are included as a reference and are expected to outperform our approach due to the additional availability of labeled target data for adaptation, we primarily focus on evaluating our performance against relevant SFDA methods, including ATCoN [28], SFTADA [4], MTRAN [11], and the concurrent work CleanAdapt [8]. On both datasets, we are able to achieve good results. For **UCF-HMDB**$_{full}$ we improve upon the previous state-of-the-art in SFDA by 1.1% (Table 1) even outperforming recent UDA works. The egocentric nature of its videos makes **EPIC-Kitchens-55** a much more difficult dataset, especially in the cross-domain scenario, yet LCMV achieves results on-par with the concurrent SFDA state-of-the-art method [8]. This emphasizes the effectiveness of using a backpropagation-free prototypical classifier, which allows to perform equally well (better on 3/6 shifts) than a network trained end-to-end on target. The proposed LCMV demonstrates good performance compared to UDA methods, being only 2.6% behind the UDA state-of-the-art (Table 2).

4.3 Class Imbalance Analysis

We provide a brief analysis of the impact of class imbalance in the **EPIC-Kitchens-55** dataset on domain shift. In fact, unlike many benchmark datasets that are artificially balanced with an equal number of samples per class, this dataset reflects the real-world scenario where class distributions are often uneven.

Table 2. Comparison with DA and SFDA methods on **EPIC-Kitchens-55**. Best UDA results are marked in **bold**, best SFDA results in **bold blue**, and the MM columns indicates which methods are multi-modal.

Method	SFDA	MM	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	Avg
DANN [10,19]	-	-	38.3	38.8	37.7	42.1	36.6	41.9	39.2
TA3N [3,26]	-	-	40.9	39.9	34.2	44.2	37.4	42.8	39.9
CoMix [19]	-	-	38.6	42.3	42.9	49.2	40.9	45.2	43.2
TransVAE [26]	-	-	**50.3**	48.0	50.5	**58.0**	**50.3**	**58.6**	**52.6**
MMD [17,29]	-	✓	46.6	39.2	43.1	48.5	48.3	55.2	46.8
MM-SADA [18]	-	✓	48.2	50.9	49.5	56.1	44.1	52.7	50.3
Kim et al. [13]	-	✓	49.5	51.5	50.3	56.3	46.3	52.0	51.0
STCDA [22]	-	✓	49.0	**52.6**	52.0	55.6	45.5	52.5	51.2
CIA [29]	-	✓	49.8	52.2	**52.5**	57.6	47.8	53.2	52.2
SHOT [15,11]	✓	✓	44.1	54.0	40.8	36.5	49.0	45.3	45.0
MTRAN [11]	✓	✓	46.3	**58.2**	42.2	38.1	**52.3**	46.1	47.2
CleanAdapt [8]	✓	✓	46.2	47.8	**52.7**	**54.4**	47.0	52.7	**50.1**
Ours	✓	✓	**52.1**	50.5	45.6	52.3	41.1	**55.1**	49.5

Table 3. Average **accuracy** and **per-class accuracy** on **EPIC-Kitchens-55**.

Method	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	Avg	
Source-only	46.3	46.8	47.6	55.3	43.0	50.3	48.2	average
Ours	52.1	50.5	45.6	52.3	41.1	55.1	49.5	average
Target-only	64.6	64.6	76.4	76.4	72.9	72.9	71.3	average
Source-only	35.9	33.4	37.2	53.8	23.6	36.8	36.8	per-class
Ours	59.8	53.2	57.0	63.7	40.6	47.5	53.6	per-class
Target-only	68.0	68.0	79.8	79.8	55.8	55.8	67.9	per-class

To demonstrate the influence of class imbalance on performance, we compare the average accuracy with the average per-class accuracy in Table 3. The average accuracy is 3.4% higher than the average per-class accuracy in the supervised setting (*Target-only*), indicating that the model is slightly better at classifying samples from common classes. In the source-only case, this discrepancy is radically increased to 11.4%, with the model achieving 0% class accuracy in some extreme cases. This shows that the model is much less robust with respect to underrepresented classes, especially in cross-domain scenarios. Our proposed LCMV does not explicitly encode any class bias, leading to an average per-class accuracy 4.1% higher than the average accuracy, and showing a significant improvement over the source-only baseline.

4.4 Multi-Modal Learning

Figure 3 shows a comparison in terms of accuracy on **UCF-HMDB**$_{full}$ using our method, source-only, target-only, and the multi-modal RGB and Flow approaches, as well as the individual modalities.

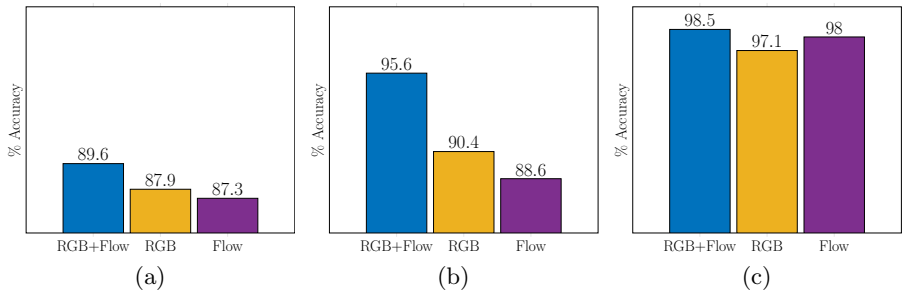

Fig. 3. Uni-modal and **multi-modal** accuracy (%) on **UCF-HMDB**$_{full}$ for **source-only** (a), **ours** (b) and **target-only** (c).

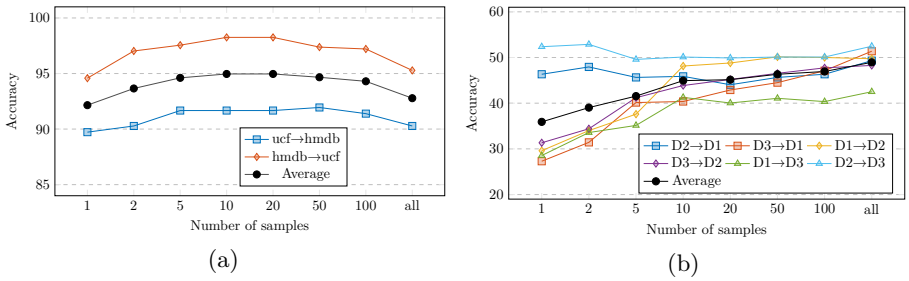

Fig. 4. Accuracy (%) on **UCF-HMDB**$_{full}$ (a) and **EPIC-Kitchens-55** (b) depending of the number t of lowest entropy samples used to compute each prototype.

Results show that training and evaluating the models on both modalities yield the best performance across all experiments. This indicates the models' ability to leverage the complementary information provided by the two modalities. While the improvement in the target-only case is relatively minor, it becomes more significant in the cross-domain scenario (source-only). Notably, our method shows a substantial improvement when using multi-modal data, achieving an increase of over 5% compared to the RGB-only baseline. This highlights the additional benefits that LCMV gains from incorporating multiple modalities.

4.5 Ablation study

In this section, we perform an ablation study to analyze the impact of key parameters and design choices in our method. Specifically, we examine the accuracy based on the number of samples used for computing each prototype and the iterative refinement of prototypes. Finally, we add some considerations on memory and time complexity.

Sample Selection. Figure 4 shows an analysis on the number of samples used for computing prototypes. Selecting the 20 lowest entropy features from each class noticeably improves our method's accuracy on **UCF-HMDB**$_{full}$. On the more difficult **EPIC-Kitchens-55**, however, we do not observe this

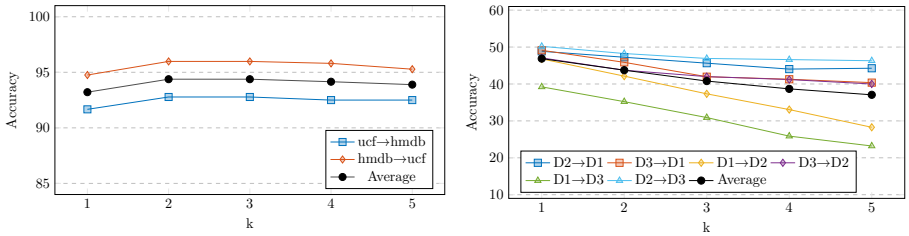

Fig. 5. Accuracy (%) on **UCF-HMDB**$_{full}$ (a) and **EPIC-Kitchens-55** (b) based on the number of iterations for prototype refinement.

pattern, with accuracy increasing proportionally to the number of features used to compute the prototypes.

Prototypes Refinement. Figure 5 shows that iteratively refining the prototypes works well on **UCF-HMDB**$_{full}$ for the first 2-3 iterations, improving performance by over 1.1%, but deteriorating thereafter. On the other hand, on **EPIC-Kitchens-55** this technique leads to gradually decreasing performance.

Memory and Time Complexity. Our backpropagation-free method is significantly more efficient than previous SFDA methods [11], requiring only 1.5GB of GPU memory (4.5 times less) and exhibiting an average execution time of 2 minutes and 33 seconds for **UCF-HMDB**$_{full}$ (3.6x faster) and 12 minutes and 20 seconds for **EPIC-Kitchens-55** (2.1x faster).

5 Conclusion

This work proposes a Source-Free Domain Adaptation method which consists in replacing the final classification layer with a prototype-based classifier. We show that the proposed LCMV improves cross-domain performance without modifying the network parameters themselves, making it more efficient than traditional backpropagation-based adaptation methods. Experiments show that LCMV is able to outperform all previous SFDA methods on two of the most important datasets for action recognition, achieving competitive results w.r.t. UDA methods. Moreover, we show that LCMV does not suffer from class imbalance, allowing for higher performance.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)

2. Chen, C.F.R., Panda, R., Ramakrishnan, K., Feris, R., Cohn, J., Oliva, A., Fan, Q.: Deep analysis of cnn-based spatio-temporal representations for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6165–6175 (2021)
3. Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., Zheng, J.: Temporal attentive alignment for large-scale video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6321–6330 (2019)
4. Chen, P., Ma, A.J.: Source-free temporal attentive domain adaptation for video action recognition. In: Proceedings of the 2022 International Conference on Multimedia Retrieval. pp. 489–497 (2022)
5. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. In: International Conference on Learning Representations (2019)
6. Choi, J., Sharma, G., Schulter, S., Huang, J.B.: Shuffle and attend: Video domain adaptation. In: Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XII 16. pp. 678–695. Springer (2020)
7. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 720–736 (2018)
8. Dasgupta, A., Jawahar, C., Alahari, K.: Overcoming label noise for source-free unsupervised video domain adaptation. In: ICVGIP 2022-Indian Conference on Computer Vision, Graphics and Image Processing. pp. 1–9. ACM (2022)
9. Fan, Q., Chen, C.F.R., Kuehne, H., Pistoia, M., Cox, D.: More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. Advances in Neural Information Processing Systems **32**, 2261–2270 (2019)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
11. Huang, Y., Yang, X., Zhang, J., Xu, C.: Relative alignment network for source-free multimodal video domain adaptation. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1652–1660 (2022)
12. Iwasawa, Y., Matsuo, Y.: Test-time classifier adjustment module for model-agnostic domain generalization. Advances in Neural Information Processing Systems **34**, 2427–2440 (2021)
13. Kim, D., Tsai, Y.H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., Chandraker, M.: Learning cross-modal contrastive features for video domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13618–13627 (2021)
14. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
15. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 6028–6039. PMLR (2020)
16. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7083–7093 (2019)
17. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105. PMLR (2015)

18. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 122–132 (2020)
19. Sahoo, A., Shah, R., Panda, R., Saenko, K., Das, A.: Contrast and mix: Temporal contrastive video domain adaptation with background mixing. Advances in Neural Information Processing Systems **34**, 23386–23400 (2021)
20. Singh, S., Arora, C., Jawahar, C.: First person action recognition using deep learned descriptors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2620–2628 (2016)
21. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
22. Song, X., Zhao, S., Yang, J., Yue, H., Xu, P., Hu, R., Chai, H.: Spatio-temporal contrastive domain adaptation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9787–9795 (2021)
23. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
24. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
25. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
26. Wei, P., Kong, L., Qu, X., Yin, X., Xu, Z., Jiang, J., Ma, Z.: Unsupervised video domain adaptation: A disentanglement perspective. arXiv preprint arXiv:2208.07365 (2022)
27. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 284–293 (2019)
28. Xu, Y., Yang, J., Cao, H., Wu, K., Wu, M., Chen, Z.: Source-free video domain adaptation by learning temporal consistency for action recognition. In: Computer Vision–ECCV 2022: 17th European Conference. pp. 147–164. Springer (2022)
29. Yang, L., Huang, Y., Sugano, Y., Sato, Y.: Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14722–14732 (2022)
30. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l 1 optical flow. In: Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29. pp. 214–223. Springer (2007)
31. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
32. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European conference on computer vision (ECCV). pp. 803–818 (2018)