A Framework for Real-World Egocentric Action Recognition

Chiara Plizzari [*,1], **Mirco Planamente** [*,1,2,3], **Emanuele Alberti**[,1], **Barbara Caputo**[,1,2]

[1]Politecnico di Torino, [2]Consortium CINI, [3]Istituto Italiano di Tecnologia (IIT)

chiara.plizzari@polito.it, mirco.planamente@polito.it,
emanuele.alberti@polito.it, barbara.caputo@polito.it

Abstract

First person action recognition is becoming an increasingly researched area thanks to the rising popularity of wearable cameras. This is bringing to light cross-domain issues that are yet to be addressed in this context. Indeed, the information extracted from learned representations suffers from an intrinsic "environmental bias". This strongly affects the ability to generalize to unseen scenarios, limiting the application of current methods to real settings where labeled data are not available during training. In this work, we propose a framework for Unsupervised Domain Adaptation (UDA) in First Person Action Recognition. To tackle the domain-shift which exists under the UDA setting, we first exploited a recent Domain Generalization (DG) technique, called Relative Norm Alignment (RNA). It consists in designing a model able to generalize well to any unseen domain, regardless of the possibility to access target data at training time. Then, in a second phase, we extended the approach to work on unlabelled target data, allowing the model to adapt to the target distribution in an unsupervised fashion. For this purpose, we included in our framework existing UDA algorithms, such as Temporal Attentive Adversarial Adaptation Network (TA^{3}N), jointly with new multi-stream consistency losses, namely Temporal Hard Norm Alignment (T-HNA) and Min-Entropy Consistency (MEC). Our approach leads to strong results in DG and UDA settings on the EPIC-Kitchens-100. Furthermore, we participate at the EPIC-Kitchens-100 Unsupervised Domain Adaptation (UDA) Challenge[1] in Action Recognition (entry 'plnet') achieving the 1st position for *'verb'*, and the 3rd position for both *'noun'* and *'action'*.

1 Introduction

First person action recognition offers a wide range of opportunities which arise from the use of wearable devices. In fact, since it intrinsically comes with rich sound information, due to the strong hand-object interactions and the closeness of the sensors to the sound source, it encourages the use of auditory information. Moreover, the continuous movement of the camera, which moves around with the observer, strongly motivates the use of secondary modalities capturing the motion in the scene, such as optical flow.

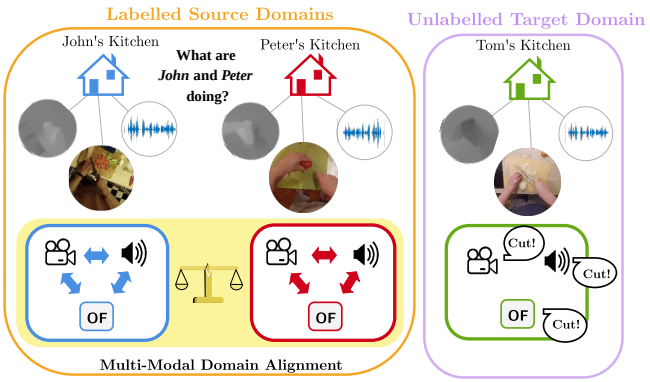

Figura 1: The correlation between the distinctive sound of an action and its corresponding visual information or motion is not always guaranteed across different domains. Thus, effectively combining multi-modal information from *multiple sources* is fundamental to increase the capability to recognize daily actions.

Our idea is that exploiting the intrinsic peculiarities of all these modalities is of crucial importance, especially in cross-domain scenarios. In fact, these modalities suffer from a domain shift which is not of the same nature. For instance, the optical flow modality, by focusing on the motion in the scene rather than on the appearance, is less sensitive to environmental changes, and thus potentially more robust than the visual modality when changing environment [Munro e Damen, 2020] (Figure 1). On the other side, the domain shift of auditory information is very different from the visual one (e.g., the sound of 'cut' will differ from a plastic to a wooden cutting board). For all those reasons, the classifier should be able to measure and understand which modality is informative and should rely on in the final prediction, and which is not.

To this purpose, authors of [Planamente *et al.*, 2022] recently proposed a multi-modal framework, called Relative Norm Alignment network (RNA-Net), which aims to progressive-

*The authors equally contributed to this work

[1]https://competitions.codalab.org/competitions/26096#results

ly align the feature norms of audio and visual (RGB) modalities among multiple sources in a Domain Generalization (DG) setting, where target data are not available during training. In that work, they bring to light that *simply feeding all the source domains to the network without applying any adaptive techniques leads to sub-optimal performance. Indeed, a multi-source domain alignment allows the network to promote domain-agnostic features.*

Interestingly, the availability of multiple sources in the official challenge dataset make it perfect to tackle the problem under a DG setting. To this purpose, we extended RNA-Net to the Flow modality, obtaining remarkable results without accessing target data. In a second stage, we further adapted it to work with unlabelled target data under the standard Unsupervised Domain Adaptation (UDA) setting. Finally, our final submission was obtained by ensembling different model streams by means of DA-based consistency losses, namely Temporal Hard Norm Alignment (T-HNA) and Min-Entropy Consistency (MEC).

2 Our Approach

In this section, we first describe the DG approach we used. Then, we illustrate its extension to unlabelled target data under the standard UDA framework. Finally, we repurpose existing DA-based losses to induce consistency between different architectures.

2.1 Domain Generalization

The multi-source nature of the proposed challenge setting makes it perfect to deal with the domain shift using DG techniques. Thus, we first exploited a method which has been recently proposed to operate in this context, called Relative Norm Alignment (RNA) [Planamente *et al.*, 2022]. This methods consists in performing an *audio-visual domain alignment* at feature-level by minimizing a cross-modal loss function ($\mathcal{L}_{RNA}$). The latter aims at minimizing the *mean-feature-norm distance* between the audio and visual features norms among all the source domains, and it is defined as

$$\mathcal{L}_{RNA} = \left(\frac{[h(X^v)]}{[h(X^a)]} - 1 \right)^2, \tag{1}$$

where $h(x_i^m) = (\|\cdot\|_2 \circ f^m)(x_i^m)$ indicates the L_2-norm of the features f^m of the m-th modality, $[h(X^m)] = \frac{1}{N}\sum_{x_i^m \in \mathcal{X}^m} h(x_i^m)$ for the m-th modality and N denotes the number of samples of the set $\mathcal{X}^m = \{x_1^m, ..., x_N^m\}$.

Authors of [Planamente *et al.*, 2022] proved that the norm unbalance between different modalities might cause the model to be biased towards the source domain that generate features with greater norm and thus causing a wrong prediction. Indeed, by simultaneously solving the problem of classification and relative norm alignment on different domains, the network extracts a shared knowledge between the different sources, resulting in a domain-agnostic model.

In our submission to the EPIC-Kitchen UDA challenge, we extended the RNA-Net framework to the optical flow modality, and we exploited the multiple sources available from the official training splits to show the effectiveness of RNA loss in a multi-source DG setting.

2.2 Domain Adaptation

In this section, we describe the UDA techniques that are integrated in our approach.

Relative Norm Alignment Network. We followed the extension towards the UDA setting proposed in [Planamente *et al.*, 2022], which is possible thanks to the unsupervised nature of RNA. In order to consider the contribution of both source and target data during training, we redefined $\mathcal{L}_{RNA}$ under the UDA setting as

$$\mathcal{L}_{RNA} = \mathcal{L}_{RNA}^s + \mathcal{L}_{RNA}^t, \tag{2}$$

where $\mathcal{L}_{RNA}^s$ and $\mathcal{L}_{RNA}^t$ correspond to the RNA formulation in Equation 1 illustrated above, when applied to source and target data respectively.

Temporal Attentive Adversarial Adaptation Network (TA^{3}N). Authors of [Chen *et al.*, 2019] proposed an UDA technique based on three components. The first one, called *Temporal Adversarial Adaptation Network (TA^{2}N)*, consists in an extension of DANN [Ganin e Lempitsky, 2015], aiming to align the temporal features on a multi-scale Temporal Relation Module (TRM) [Zhou *et al.*, 2018] through a gradient reversal layer (GRL). The second component is based on a domain attention mechanism which guides the temporal alignment towards features where the domain discrepancy is larger. Finally, the third component uses a minimum entropy regularization (attentive entropy) to refine the classifier adaptation.

2.3 Ensemble UDA losses

For our final submission, different models are used in order to exploit the potentiality of popular video architectures. Training individually each backbone with standard UDA protocols results in an adapted feature representation which varies from stream to stream. Our intuition is that this aspect could impact negatively the training process and the performance on target data. In fact, since the domain adaption process acts on each architecture independently, different prediction logits are obtained on target data. When combining them, this could cause a mismatch between the final scores, increasing the level of uncertainty of the model. Thus, we impose a consistency constraint between feature representations from different models, by repurposing existing UDA loss functions to operate between multiple streams. Those are:

Temporal Hard Norm Alignment (T-HNA). It rebalances the contribution of each model during training by extending HNA [Planamente *et al.*, 2022] to align the norms of features coming from the different streams towards the same value R. This is applied on features extracted from multiple scales of each TRN module. The resulting $\mathcal{L}_{T\text{-}HNA}$ is defined as

$$\mathcal{L}_{T\text{-}HNA} = \sum_b \left([h_t(X^b)] - R \right)^2, \tag{3}$$

where h_t denotes the L_2-norm of features extracted from the t-th multi-scale level of the b-th backbone network.

Min Entropy Consensus (MEC loss). We extended the loss proposed in [Roy *et al.*, 2019] to encourage coherent

Unsupervised Domain Adaptation Leaderboard							
	Rank	Verb Top-1	Noun Top-1	Action Top-1	Verb Top-5	Noun Top-5	Action Top-5
chengyi	1	53.16	34.86	**25.00**	80.74	59.30	40.75
M3EM	2	53.29	**35.64**	24.76	81.64	59.89	40.73
plnet	3	**55.22**	34.83	24.71	**81.93**	**60.48**	**41.41**
EPIC_TA3N [Damen *et al.*, 2020]	6	46.91	27.69	18.95	72.70	50.72	30.53
EPIC_TA3N_SOURCE_ONLY [Damen *et al.*, 2020]	12	44.39	25.30	16.79	69.69	48.40	29.06

Tabella 1: Leaderboard results of EPIC-Kitchens Unsupervised Domain Adaptation Challenge. The results obtained by the top-3 participants and the provided baseline methods are reported. **Bold:** highest result; **Green:** our final submission.

Ensemble UDA losses						
	Top-1			Top-5		
	Verb	Noun	Action	Verb	Noun	Action
Ensemble	52.83	30.82	21.96	**81.04**	52.67	46.66
Ensemble+T-HNA	53.84	32.54	22.65	80.63	54.86	48.03
Ensemble+T-HNA+MEC	**54.02**	**33.53**	**23.58**	81.00	**55.03**	**48.27**

Domain Generalization			
	Target	Verb Top-1	Verb Top-5
Source Only	✗	44.39	69.69
EPIC_TA3N [Damen *et al.*, 2020]	✓	46.91	72.70
RNA-Net [Planamente *et al.*, 2022]	✗	47.96	79.54
EPIC_TA3N+RNA-Net	✓	**50.40**	**80.47**

Tabella 2: **Left.** Results on the EPIC-Kitchen validation set with different ensembling UDA losses. **Right.** Results on EPIC-Kitchen test set under the DG setting. **Bold** highest result.

predictions between different models. The resulting loss is defined as:

$$\mathcal{L}_{MEC} = -\frac{1}{m}\sum_{i=1}^{m}\frac{1}{b}\max_{y\in\mathcal{Y}}\sum_{b}\log p_b(y|x_i^t) \qquad (4)$$

where m is the cardinality of the batch size of the target set, y is the predicted class, and $\log p_b(y|x_i^t)$ is the prediction probability of the b-th backbone network. The intuitive idea behind the proposed approach is to encourage different backbones to have a similar predictions.

3 Framework

In this section, we describe the architectures of the feature extractors used to produce suitable multi-modal video embeddings, and the fusion stategies adopted to combine them. We complete this section with the description of the hyper-parameters used for the training.

3.1 Architecture

Backbone. For our submission, we adopted different network configurations. In the first one, corresponding to the RNA-Net framework in [Planamente *et al.*, 2022], we used the Inflated 3D ConvNet (I3D), pre-trained on Kinetics [Carreira e Zisserman, 2017], for RGB and Flow streams, and a BN-Inception model [Ioffe e Szegedy, 2015] pre-trained on ImageNet [Deng *et al.*, 2009] for the auditory information. Each feature extractor produces a 1024-dimensional representation which is fed to an action classifier. In the second configuration, we used BNInception for all the three streams, using pre-extracted features from a TBN [Munro e Damen, 2020] model trained on EPIC-Kitchens-55. In the last configurations, we used standard ResNet50 [He *et al.*, 2016] for all the streams using TSN [Wang *et al.*, 2016] and TSM [Lin *et al.*, 2019] models pre-trained on Epic-Kitchen55[2].

[2]https://github.com/epic-kitchens/epic-kitchens-55-action-models

λ_{RNA}	λ_{HNA}	R	λ_{MEC}	γ	β
1	0.0006	40	0.01	0.003	0.75, 0.75, 0.5

Tabella 3: UDA losses hyper-parameters used during training.

Multi-modal fusion strategies. In all the above mentioned configurations, each modality is processed by its own backbone, and the corresponding extracted representations are then fused following different strategies. For RNA-Net, we followed a standard late fusion strategy, consisting in averaging the final score predictions obtained from two different fully-connected layers (verb, noun) from each modality. In the other configurations, we adopted the mid-fusion strategy proposed in [Kazakos *et al.*, 2019], to generate a common frame-embedding among the modalities and used a Temporal Relation Module (TRM) [Zhou *et al.*, 2018] to aggregate features from different frames before feeding the final embeddings to the verb and noun classifiers.

3.2 Implementation Details

We trained I3D and BNInception models with SGD optimizer, with an initial learning rate of 0.001, dropout 0.7, and using a batch size of 128, following [Planamente *et al.*, 2022]. Instead, when using pre-extracted features from ResNet50 or BNInception, we trained the TRM modules on top of them for 100 epochs with an initial learning rate of 0.03, decayed after epochs 30 and 60 by a factor of 0.1. We used a batch size of 128 with SGD optimizer. In Table 3 we report the other hyper-parameter used. Specifically, we indicate with λ_{RNA}, λ_{T-HNA} and λ_{MEC} the weights of RNA, T-HNA and MEC losses respectively, and with R the values of the radius of T-HNA (see Equation 4). In addition, we report the values used in TA^{3}N to weight the attentive entropy loss (γ) and the domain losses at different levels (β).

4 Results and Discussion

In Table 1 we report our best performing model on the target test, achieving the **1st** position on 'verb', **3rd** on 'noun' and 'action', and **1st** position on Top-5 accuracy on all categories. In Table 2 (left) we show an ablation on the contribution of the proposed ensemble UDA losses, T-HNA and MEC respectively, on the official validation set. As it can be seen, they improve Top-1 accuracy on all categories by up to 2%, proving the effectiveness of imposing a consistency between features from different streams.

How well do DG approaches perform? We show in Table 2 (right) the results obtained under the multi-source DG setting, when target data are not available during training. Noticeably, RNA outperforms the baseline Source Only by up to 3% on Top-1 and 10% on Top-5, remarking the importance of using ad-hoc alignment techniques to deal with multiple sources in order to effectively extract a domain-agnostic model. Moreover, it outperforms the very recent UDA technique TA^{3}N without accessing to target data. Interestingly, when combined with EPIC_TA3N, it further improves performance, proving the complementarity of RNA to other existing UDA approaches.

Riferimenti bibliografici

[Carreira e Zisserman, 2017] Joao Carreira e Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[Chen *et al.*, 2019] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, e Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6321–6330, 2019. 2

[Damen *et al.*, 2020] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 3

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, e Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[Ganin e Lempitsky, 2015] Yaroslav Ganin e Victor Lempitsky. Unsupervised domain adaptation by backpropagation. volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR. 2

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, e Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[Ioffe e Szegedy, 2015] Sergey Ioffe e Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach e David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456. PMLR, 2015. 3

[Kazakos *et al.*, 2019] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, e Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3

[Lin *et al.*, 2019] Ji Lin, Chuang Gan, e Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019. 3

[Munro e Damen, 2020] Jonathan Munro e Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020. 1, 3

[Planamente *et al.*, 2022] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, e Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1807–1818, 2022. 1, 2, 3

[Roy *et al.*, 2019] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, e Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9471–9480, 2019. 2

[Wang *et al.*, 2016] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, e Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 3

[Zhou *et al.*, 2018] Bolei Zhou, Alex Andonian, Aude Oliva, e Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 2, 3