POLITECNICO DI TORINO
Repository ISTITUZIONALE

PoliTO-IIT-CINI Submission to the EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge for Action Recognition

Original
PoliTO-IIT-CINI Submission to the EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge for Action Recognition / Planamente, Mirco; Goletto, Gabriele; Trivigno, Gabriele; Averta, Giuseppe; Caputo, Barbara. - (2022). (Intervento presentato al convegno The Tenth International Workshop on Egocentric Perception, Interaction and Computing (WCVPR22)).

Availability:
This version is available at: 11583/2971271 since: 2022-09-13T17:59:33Z

Publisher:
EPIC-KITCHENS

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

24 September 2024

arXiv:2209.04525v1 [cs.CV] 9 Sep 2022

PoliTO-IIT-CINI Submission to the EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge for Action Recognition

Mirco Planamente[1,2,3] Gabriele Goletto [1] Gabriele Trivigno [1] Giuseppe Averta [1] Barbara Caputo[1,3]

[1] Politecnico di Torino
name.surname@polito.it

[2] Istituto Italiano di Tecnologia
name.surname@iit.it

[2] Consortium Cini

Abstract

In this report, we describe the technical details of our submission to the EPIC-Kitchens-100 Unsupervised Domain Adaptation (UDA) Challenge in 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). Secondly, we extended this approach to work on unlabelled target data, enabling a simpler adaptation of the model to the target distribution in an unsupervised fashion. To this purpose, we included in our framework UDA algorithms, such as multi-level adversarial alignment and attentive entropy. By analyzing the challenge setting, we notice the presence of a secondary concurrence shift in the data, which is usually called environmental bias. It is caused by the existence of different environments, i.e., kitchens. To deal with these two shifts (environmental and temporal), we extended our system to perform Multi-Source Multi-Target Domain Adaptation. Finally, we employed distinct models in our final proposal to leverage the potential of popular video architectures, and we introduced two more losses for the ensemble adaptation. Our submission (entry 'plnet') is visible on the leaderboard and ranked in 2nd position for 'verb', and in 3rd position for both 'noun' and 'action'.

1. Introduction

First person action recognition offers a wide range of opportunities and challenges, thanks to the use of wearable devices to capture the current state of the user and of the environment. Very often, indeed, the actions of the subject are captured through a video-camera placed on the head of the user. As a consequence, in contrast with most CV tasks, the major feature of this scenario is that source data is intrinsically characterized by rich multi-modal information, thanks to the proximity of the sensor to the action scene. As a result, sensor fusion between visual and auditory cues can be a powerful method to fully exploit the knowledge available in the data. However, the particular setup of data collection also comes with several difficulties: i) ego-motion represents a significant source of noise for the dataset, because changes in head posture cause a shift in the point-of-view and background. While from one side this effect can be exploited as an intrinsic attention mechanism, it may also introduce confusion between ego-motion and the real action of the subject. An approach to mitigate this effect could be to complement RGB data with other motion-related sources, such as the optical flow; ii) model predictions tend to be strongly correlated with the surrounding environment, which represents a bias in the dataset (usually referred to as *environmental bias*), thus resulting in decreased performances when the environment changes (e.g. different kitchens). In this report, we discuss the idea that, to fully exploit the potential of data sources, and to mitigate the performances drop across domains, it is crucial to properly combine several sensing modalities, including audio, video, and motion. This is particularly true for cross-domain scenarios, where test data are extracted from a different distribution w.r.t. the training data (i.e. different users and/or kitchens). Indeed, the effect of domain shift is not consistent across different sensing modalities, and some of them may suffer in some cases where others are more robust.

The reason is that domain shifts are not all of the same nature. For instance, the optical flow is more focused on the motion in the scene, rather than appearance, and is therefore less sensitive to environmental changes, thus showing higher robustness than the visual modality when changing environment [12]. 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 assess - depending on the conditions - which modality is more informative, and therefore should be considered more for the final prediction.

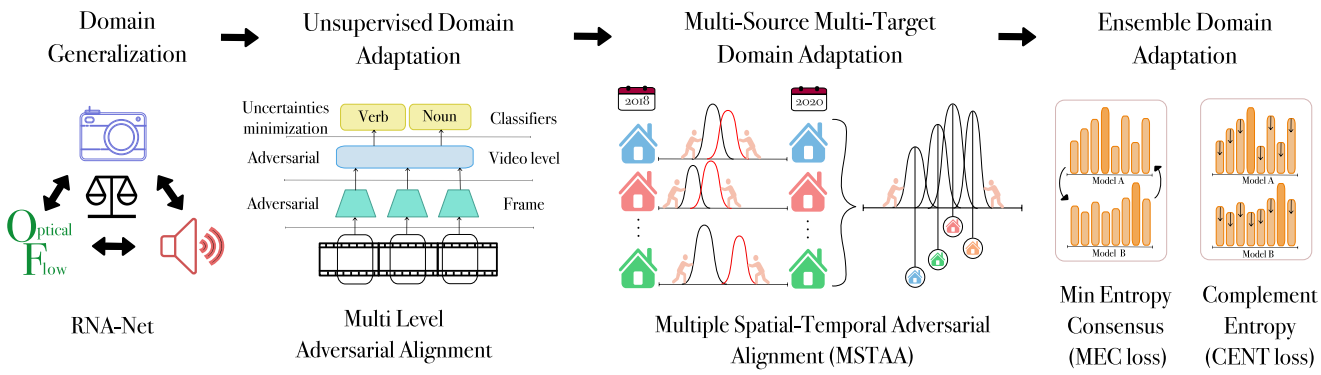

Figure 1: An overview of the proposed approach. It can be summarized in four main aspects: **1.** Domain Generalization through RNA-Net [15], **2.** Unsupervised Domain Adaptation via Multi-Level Adversarial Alignment and entropy minimization, **3.** Multi-Source Multi-Target Domain Adaptation extension and **4.** Ensemble Domain Adaptation losses.

To this purpose, authors of [15] recently proposed a multi-modal framework, called Relative Norm Alignment network (RNA-Net), which aims at progressively aligning 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. Interestingly, the authors showed that *merely feeding all the source domains to the network without applying any adaptive techniques leads to sub-optimal performance, while a multi-source domain alignment allows the network to promote domain-agnostic features.*

Including all the aforementioned considerations, we developed the method adopted in the challenge with the following steps (see also Figure 1):

1. RNA-Net was extended to the Flow modality, obtaining remarkable results without accessing target data;
2. with further modifications, RNA-Net was adapted to work with unlabelled target data under the standard Unsupervised Domain Adaptation (UDA) setting;
3. the challenge's setting was revisited by identifying a new concurrent shift denominated "environmental bias". Our framework was modified accordingly to perform Multi-Source Multi-Target Domain Adaptation;
4. the final submission was obtained by combining different model streams by means of DA-based losses, namely Min-Entropy Consistency (MEC) and Complement Entropy (CENT).

2. Our Approach

In this section, we first describe the DG approach used. Then, we show our UDA framework and its extension for Multi-Source Multi-Target Domain Adaptation. Finally, we demonstrate how to re-define 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) [15]. This methods consists of an *audio-visual domain alignment* at feature-level through the minimization of 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

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

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, $\mathbb{E}[h(X^m)] = \frac{1}{N} \sum_{x_i^m \in \mathcal{X}^m} h(x_i^m)$ 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 [15] 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, thus causing wrong predictions. Contrarily, 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, in order to exploit the multiple sources available from the official training splits while showing the effectiveness of RNA loss in a multi-source DG setting.

2.2. Domain Adaptation

The UDA techniques embedded into our pipeline can be divided in two main groups: *feature*-level and *classifier*-level. The first aims at aligning the distribution of source and target, and works at different levels of representation (frames- and video-level); the latter, instead, reduces the classifier's uncertainty on target data.

Multi-Level Adversarial Alignment.

Following popular practices in unsupervised video domain adaption techniques, we integrate into our framework an adversarial approach [3, 12], consisting of an extension of the DANN [8] standard UDA image-based method. We apply it at two different feature levels; frame- and video-level. It entails the introduction of two separate branches in our framework. Down-stream of said branches there are discriminators that try to distinguish the two domains (source and target). Contrarily, by maximising the corresponding discriminator losses, the network learns feature representations invariant to both domains.

Attentive Entropy. In order to reduce the uncertainty of the classifier on the target data, we minimize the attentive entropy loss proposed in [3] as in [17]. This action minimizes the entropy, resulting in a refinement of the classifier adaptation. The term "attentive" refers to a loss reweighting approach that prioritizes videos with low domain discrepancy by focusing on minimizing entropy for these videos.

2.3. Multi-Source Multi-Target Domain Adaptation

The previous Epic Kitchen challenges [6, 5], as well as the literature on unsupervised domain adaptation for first person action recognition [13, 15, 14, 18, 16], reveal a strong dependency of the models on the environment where the actions are recorded. This problem, known as "environmental bias", causes a decrease in performance in occurrence of environment switches. As regards past action recognition challenges, we see this behavior by comparing performances of the models when tested on S1 (seen) and S2 (unseen). In the setting proposed in [13], similar behavior is observed, demonstrating the model's low generalization ability when tested on different kitchens.

The above considerations allow us to identify a secondary shift in this challenge, that occurs along with the temporal shift. Indeed, the training data are collected from different environments i.e. kitchens, thus introducing an environmental shift. As a result, we may rename the challenge setting *Multi-Source Multi-Target Unsupervised Domain Adaptation*.

To deal with this new setting we propose a novel framework, which we call Multiple Spatio-Temporal Adversarial Alignment (MSTAA), combining Multiple Temporal Adversarial Alignment (MTAA) and Multiple Spatial Adversarial Alignment (MSAA). MTAA is obtained by adopting 2K domain adversarial branches (where K indicates the number of kitchens), aligning the source and the target distribution both at video- and frame-level for each kitchen. Instead, MSAA consists in adding another adversarial branch with a k-dimension discriminator in order to align the distribution of different kitchens and alleviate the environmental bias issue.

2.4. Ensemble UDA losses

For our final submission different models have been used in order to fully exploit the potentiality of popular video architectures. However, training individually each backbone with standard UDA protocols would result in independently adapted feature representations, which consequently vary between different streams. Our intuition is that this aspect could impact negatively the training process and the performance on target data. Indeed, since the domain adaption process acts on each architecture independently, naively training the backbones separately would yield mismatching prediction logits on target data, which, when combined, could increase the level of uncertainty of the model. For this reason, we use the Min Entropy Consensus (MEC) loss, to impose a consistency constraint between feature representations from various models. Then, re-purposing the existing Complement Entropy (CENT) loss, we attempt to exploit the target data samples based on the assumption that there are some conditions in which it is easier to answer the question *"Which classes does this action not belong to?"* rather than *"Which class does this action belong to?"*.

Min Entropy Consensus (MEC loss). We extended the loss proposed in [19] to encourage coherent predictions between different models. The resulting loss is defined as:

$$\mathcal{L}_{MEC} = -\frac{1}{m}\sum_{i=1}^{m}\frac{1}{b}\max_{y\in\mathcal{Y}}\sum_{b}\log p_b(y|x_i^t) \qquad (2)$$

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.

Complement Entropy (CENT). The Complement Entropy (CENT) loss aims at neutralizing the negative effects on the final prediction of clips whose logits present high degrees of uncertainty. It accomplishes this by "flattening" the predicted probabilities of "complement classes", i.e., all classes except the predicted one. As a result, when predictions are ensembled, the noise due to uncertainty on complement classes is reduced. We refer to this loss as "complement entropy" objective, as it consists in maximizing the entropy for low-confident classes rather than minimizing it for the most confident one, as standard entropy minimization does. It is defined as:

UNSUPERVISED DOMAIN ADAPTATION LEADERBOARD							
	Rank	Verb Top-1	Noun Top-1	Action Top-1	Verb Top-5	Noun Top-5	Action Top-5
VI-I2R	1	**57.89**	40.07	**30.12**	**83.48**	64.19	**48.10**
Audio-Adaptive-CVPR2022	2	52.95	**42.26**	28.06	80.03	**67.51**	44.03
plnet	3	55.51	35.86	25.25	82.77	60.65	40.09
CVPR2021-chengyi	4	53.16	34.86	25.00	80.74	59.30	40.75
CVPR2021-M3EM	5	53.29	35.64	24.76	81.64	59.89	40.73
CVPR2021-plnet	6	55.22	34.83	24.71	81.93	60.48	41.41
EPIC_TA3N [4]	8	46.91	27.69	18.95	72.70	50.72	30.53
EPIC_TA3N_SOURCE_ONLY [4]	9	44.39	25.30	16.79	69.69	48.40	29.06

Table 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 Underline: second highest result; **Green:** our final submission.

UNSUPERVISED DOMAIN ADAPTATION			
	Verb	Noun	Action
Ensemble (E) *Source Only*	53.64	32.65	22.98
E-UDA	53.88	33.10	23.22
E+MEC	53.67	34.32	23.91
E+MEC+CENT	54.20	33.92	23.99
E-SMR+MEC+CENT	**54.55**	34.72	24.22
E-SMR+MEC+CENT+MTAA	54.09	33.72	23.77
E-SMR+MEC+CENT+MSTAA	54.01	**34.82**	**24.24**

Table 2: Results on the EPIC-Kitchen validation set.

DOMAIN GENERALIZATION			
	Target	Verb Top-1	Verb Top-5
Source Only	✗	44.39	69.69
EPIC_TA3N [4]	✓	46.91	72.70
RNA-Net [15]	✗	47.96	79.54
EPIC_TA3N+RNA-Net	✓	**50.40**	**80.47**

Table 3: Results on the EPIC-Kitchen test set.

$$\begin{aligned}\mathcal{L}_{CENT} &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(\hat{y}_{i\bar{c}}) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,j\neq p}^{\mathcal{C}}\left(\frac{\hat{y}_{ij}}{1-\hat{y}_{ip}}\log\frac{\hat{y}_{ij}}{1-\hat{y}_{ip}}\right)\end{aligned} \tag{3}$$

where N is the total number of samples in the batch, $\hat{y}_{ip}$ represents the predicted probability of the class p with the higher score for the i-th sample, i.e., $\hat{y}_{ip} = max_j(\hat{y}_{ij})$, and $\mathcal{H}(\cdot)$ is the entropy function computed on the prediction of complement classes $\hat{y}_{i\bar{c}}$ ($\bar{c} \neq p$). The formulation is similar to the one in [2], and we extend it to operate in an unsupervised fashion.

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. Finally, we deepen the analysis describing the hyper-parameters used for the training.

3.1. Architecture

Backbone. For our submission, we adopted three different network configurations. In the first one, corresponding to the RNA-Net framework in [15], we used the Inflated 3D ConvNet (I3D), pre-trained on Kinetics [1], for RGB and Flow streams, and a BN-Inception model [10] pre-trained on ImageNet [7] 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 BN-Inception models for all the three streams, using pre-extracted features from a TBN [12] model trained on EPIC-Kitchens-55. In the last configurations, we used standard ResNet-50 architectures [9] equipped with the Temporal Shift Module [11] pre-trained on EPIC-Kitchens-55 [1].

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 recent mid-fusion strategy, called Semantic Mutual Refinement submodule (SMR), proposed in [20], to generate a common frame-embedding among the modalities. Then, using tem-

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

λ_{RNA}	λ_{CENT}	λ_{MEC}	γ	β
1	0.31	0.22	0.003	0.75, 0.75, 0.75

Table 4: UDA losses hyper-parameters used during training.

poral pooling, we obtain a final video-embedding that is sent to the verb and noun classifiers.

3.2. Implementation Details

We trained I3D and BN-Inception models with SGD optimizer, with an initial learning rate of 0.001, dropout 0.7, and using a batch size of 128, following [15]. Instead, when using pre-extracted features from ResNet50 or BN-Inception, we trained the SMR modules on top of them for 45 epochs with an initial learning rate of 0.03, decayed after epochs 25 and 35 by a factor of 0.1. We used a batch size of 128 with SGD optimizer. In Table 4 we report the other hyper-parameter used. Specifically, we indicate with λ_{RNA}, λ_{CENT} and λ_{MEC} the weights of RNA, CENT and MEC losses respectively. In addition, we report the values used to weight the attentive entropy loss (γ) and the domain losses at different levels (β) for MSTAA.

4. Results and Discussion

In Table 1 we report our best performing model on the target test, achieving the **2st** position on 'verb', and the **3rd** on 'noun' and 'action'. Meanwhile, in Tables 2 and 3 we show an ablation of the proposed UDA and DG methods described in section 2.

How well do DG approaches perform? The results in Table 3 are 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, highlighting 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 recent UDA technique TA^{3}N [3] without accessing target data. Interestingly, when combined with EPIC_TA3N, it further improves performance, proving the complementarity of RNA to other existing UDA approaches.

In Table 2 it can be seen how the proposed UDA approaches improve Top-1 accuracy on all categories by up to 1%. Although using an additional adversarial branch for each kitchen does not appear to provide a significant improvement on the validation set, it increases the top-1 action accuracy on the test set, allowing us to obtain the third position in the challenge. Without MSTAA, the accuracy on the action top-1 reaches just 24.83%. This outcome was predictable given that the validation set is populated with a different set of kitchens than the test set, whereas the kitchens in the test set are the same as those used for the target and source training. This aspect confirms the *Multi-Source Multi-Target Unsupervised Domain Adaptation* setting and the presence of two different shifts, the *temporal* shift (2018-2020) and the *environmental* shift (among the kitchens).

Acknowledgements. This work was supported by the CINI Consortium through the VIDESEC project and by the Italian Ministry of University and Research under the DM1061. The research herein was carried out using the IIT HPC infrastructure.

References

[1] Joao Carreira and 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.

[2] Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Complement objective training. *arXiv preprint arXiv:1903.01182*, 2019.

[3] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and 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.

[4] 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.

[5] Dima Damen, Evangelos Kazakos, Will Price, Jian Ma, and Hazel Doughty. Epic-kitchens-55 - 2020 challenges report. `https://epic-kitchens.github.io/Reports/EPIC-KITCHENS-Challenges-2020-Report.pdf`, 2020.

[6] Dima Damen, Will Price, Evangelos Kazakos, Antonino Furnari, and Giovanni Maria Farinella. Epic-kitchens - 2019 challenges report. `https://epic-kitchens.github.io/Reports/EPIC-Kitchens-Challenges-2019-Report.pdf`, 2019.

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

[8] Yaroslav Ganin and 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.

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

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and 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.

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

[12] Jonathan Munro and 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.

[13] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[14] Mirco Planamente, Andrea Bottino, and Barbara Caputo. Self-supervised joint encoding of motion and appearance for first person action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8751–8758. IEEE, 2021.

[15] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and 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.

[16] Mirco Plananamente, Chiara Plizzari, and Barbara Caputo. Test-time adaptation for egocentric action recognition. In *International Conference on Image Analysis and Processing*, pages 206–218. Springer, 2022.

[17] Chiara Plizzari, Mirco Planamente, Emanuele Alberti, and Barbara Caputo. Polito-iit submission to the epic-kitchens-100 unsupervised domain adaptation challenge for action recognition. *arXiv preprint arXiv:2107.00337*, 2021.

[18] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. Eˆ 2 (go) motion: Motion augmented event stream for egocentric action recognition. *arXiv preprint arXiv:2112.03596*, 2021.

[19] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, and 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.

[20] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Epic-kitchens-100 unsupervised domain adaptation challenge for action recognition 2021: Team m3em technical report. *arXiv preprint arXiv:2106.10026*, 2021.