POLITECNICO DI TORINO
Repository ISTITUZIONALE

Test-Time Adaptation for~Egocentric Action Recognition

Original
Test-Time Adaptation for~Egocentric Action Recognition / Planamente, Mirco; Plizzari, Chiara; Caputo, Barbara. - 13233 LNCS:(2022), pp. 206-218. (Intervento presentato al convegno International Conference on Image Analysis and Processing, ICIAP 2022 tenutosi a Lecce (IT) nel May 23–27, 2022) [10.1007/978-3-031-06433-3_18].

Availability:
This version is available at: 11583/2970227 since: 2022-07-21T17:43:22Z

Publisher:
Springer

Published
DOI:10.1007/978-3-031-06433-3_18

Terms of use:

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

Publisher copyright
Springer postprint/Author's Accepted Manuscript (book chapters)

This is a post-peer-review, pre-copyedit version of a book chapter published in Lecture Notes in Computer Science book series (LNCS,volume 13233). The final authenticated version is available online at: http://dx.doi.org/10.1007/978-3-031-06433-3_18

(Article begins on next page)

12 August 2024

Test-Time Adaptation for Egocentric Action Recognition

Mirco Plananamente*[1,2][0000−0001−7238−1867], Chiara Plizzari*[1][0000−0003−4984−7432], and Barbara Caputo[1,2][0000−0001−7169−0158]

[1]Politecnico di Torino, Torino, Italy, [2]CINI Consortium, Roma, Italy
{mirco.planamente, chiara.plizzari, barbara.caputo}@polito.it

Abstract. Egocentric action recognition is becoming an increasingly researched topic thanks to the rising popularity of wearable cameras. Despite the numerous publications in the field, the learned representations still suffers from an intrinsic "environmental bias". To address this issue, domain adaptation and generalization approaches have been proposed, which operate by either adapting the model to target data *during training* or by learning a model able to generalize to unseen videos by exploiting the knowledge from multiple source domains. In this work, we propose to adapt a model trained on source data to novel environments *at test time*, making adaptation practical to real-world scenarios where target data are not available at training time. On the popular EPIC-Kitchens dataset, we present a new benchmark for Test-Time Adaptation (TTA) in egocentric action recognition. Moreover, we propose a new multi-modal TTA approach, which we call RNA^{++}, and combine it with a new set of losses aiming at reducing classifier's uncertainty, showing remarkable results w.r.t. existing TTA methods inherited from image classification. Code available: https://github.com/EgocentricVision/RNA-TTA.

Keywords: egocentric action recognition · test-time adaptation

1 Introduction

In the last years, the technological advances in the field of wearable devices led to a growing interest in egocentric vision due to the possibility to capture information about how humans perceive the world and interact with the environment, without the need of a fixed recording system. The first person perspective unlocks a variety of applications, including wearable sport cameras, human-robot interaction, and human assistance. Contrary to traditional third-person views, the recording equipment is worn by the observer and it moves with her, posing new issues such as ego-motion, occluded objects, and significantly more variations in lighting, perspective, and environment.

The recent release of the EPIC-Kitchens large-scale dataset [8], as well as the contests that accompanied it, has sparked interest in more efficient architectures

* The authors equally contributed to this work.

capable of dealing with these issues. Despite the numerous publications in the field [32], egocentric action recognition still has one major flaw that remains unsolved, known as "environmental bias" [39]. This problem arises from the network's heavy reliance on the environment in which the activities are recorded, which inhibits the network's ability to recognize actions when they are conducted in unfamiliar (unseen) surroundings. In general, this problem is referred to in the literature as *domain shift*, meaning that a model trained on a source labeled dataset cannot generalize well on an unseen dataset, called target. Usually, it is addressed by reducing the problem to an unsupervised domain adaptation (UDA) setting [25], where an unlabeled set of samples from the target is available and used to learn and adapt the model to the target distribution.

However, the UDA scenario is not always realistic, as (i) the target domain should be known a priori and (ii) the target data should be available at training time. To overcome those limitations, authors of [29] proposed an alternative solution which simply leverages the shared knowledge from multiple sources available during training to learn a representation that is able to generalize to any unseen domain, regardless of the possibility to access target data – known as Domain Generalization (DG) setting.

Differently from previous works, in this paper we investigate a solution that focuses on performing adaptation *during testing*. The proposed approach is based on the simple assumption that the samples received by the network during testing can be considered as a hint of the target distribution. Thus, we seek to adapt a pre-trained model to new videos coming from the test set. To best of our knowledge, this approach, known as *Test-Time Adaptation* (TTA), has never been examined in an egocentric context before. Indeed, its use in this context is even more relevant, as (i) since online adaptation does not require additional parameters, it increases the portability on multiple devices and its access to diverse users and (ii) as test data is not required to be stored, it respects privacy concerns; this is of crucial importance in the case of the first person videos as anonymization is more difficult than stardard third person videos or images [38].

In this work, we present a new benchmark for multi-modal TTA in egocentric action recognition on the well-known EPIC-Kitchens dataset. Moreover, we propose a new TTA approach, called RNA^{++}, which extends RNA-Net [29], a recent multi-modal DG method, to operate on different video clips at test time. We further combine it with a new set of losses meant to reduce the classifier's confusion on test data. Results show the effectiveness of multi-modal learning in enhancing the ability of the model to adapt to new data and further validate the effectiveness of the proposed methods.

2 Related Works

2.1 Egocentric Action Recognition

The community's interest for *First Person Action Recognition* (FPAR) has quickly grown in recent years. FPAR's architectures are generally inherited from third-person literature [43,23,3]. However, due to the complexity of the setting, the

multi-modal approach is the most popular technique, consisting in combining traditional visual RGB data with motion data, such as optical flow [3,25,43,36,10,19]. However, as shown in [7], the use of optical flow limits the application of several methods in online scenarios, pushing the community either towards single-stream architectures [49,7,28], or to investigate alternative modalities, e.g., audio [20,19] or event data [31]. This work is the first one exploiting audio modality, jointly with its visual counterpart, in a test-time adaptation scenario.

2.2 Cross-Domain Action Recognition

Under the *Unsupervised Domain Adaptation (UDA)* setting, an unlabeled set of samples from the target is available for adaptation during training. Most of the approaches have been designed for image classification tasks [11,24,22,13]. Recently, many works started to analyze UDA for video classification tasks [5,25,17,27,35,21]. Those use adversarial learning with temporal attention [27,5], multi-modal cues [25], clip order prediction [6] or contrastive losses [35,21].

The *Domain Generalization (DG)* setting, instead, aims at finding a representation able to generalize to any unseen domain, regardless of the possibility to access target data at training time. Existing approaches in DG are mostly designed for image data [40,9,2]. Only one work investigated the DG setting in third person action recognition [46]. Recently, authors of [29] proposed a solution to this problem in first person action recognition, by proposing a feature-level solution which exploits the collaboration of audio-visual signals.

In this work, we further explore the possibility to adapt the model directly on test data under a *test-time adaptation* setting. While the latter has been widely explored on image data [37,42,48,26,33], only one work explored the possibility to adopt it on videos [1]. In this work, we take a step ahead by extending the setting to the egocentric action recognition scenario.

3 Problem Formulation

Test-Time Adaptation (TTA) for Action Recognition. This setting consists in learning the target distribution using just the unlabelled videos available during test. Due the capability of wearable devices to capture data in a variety of situations and surroundings, the target distribution is extremely variable and hard to generalize to using DG techniques. Moreover, the availability of a set of target data to learn the unseen distribution from during training, as well as the continuous access to source data to re-train the model on novel environments, are both impracticable in this case, making the Source Free DA and UDA settings unfeasible. In this work, we propose TTA as an intriguing and significant setting that has yet to be investigated in the egocentric literature. Indeed, it allows to optimize the network on test data during inference by introducing an additional, but negligible w.r.t. standard training, inference cost (Table 1).

Under the video setting, two aspects have to be considered, (i) *multi-modality*, which translates in a multi-modal input $x = (x_i^v, x_i^a)$, where we denote with v and

Table 1. Adaptation settings differ by the data and losses used during train and test. The terms x_s and y_s refer to the labeled distribution, known as *source*, while x_t the unlabelled one, known as *target*. Our TTA setting only needs the target data x_t.

Setting	Source	Target	Train Loss	Test Loss
Unsupervised Domain Adaptation (UDA)	x_s, y_s	x_t	$\mathcal{L}(x_s, y_s) + \mathcal{L}(x_s, x_t)$	-
Domain Generalization (DG)	x_s, y_s	-	$\mathcal{L}(x_s, y_s)$	-
Source-Free DA	-	x_t	$\mathcal{L}(x_t)$	-
Test-Time Adaptation (TTA)	x_s, y_s	x_t	-	$\mathcal{L}(x_t)$

a the visual and audio modality respectively, and with i the i-th sample, and (ii) *temporality*, consisting in having an input x_i^m composed of k *clips* representing different temporal positions within the video, i.e., $x_i^m = \{x_{i1}^m, ..., x_{ik}^m\}$.

Problem Setting. We assume a model trained on different source domains $\{\mathcal{S}_1, ..., \mathcal{S}_n\}$, where each $\mathcal{S} = \{(x_{s,i}, y_{s,i})\}_{i=1}^{N_s}$ is composed of N_s source samples with label space Y_s known, and a target domain $\mathcal{T} = \{x_{t,i}\}_{i=1}^{N_t}$ of N_t target samples whose label space Y_t is unknown. The main assumptions is that the label space is shared, $\mathcal{Y}_s = \mathcal{Y}_t$. Our objective is to perform *test-time adaptation* by adapting the model trained on source data to samples available at test time. During the forward pass, each modality input (x_i^v, x_i^a) is fed to a separate feature extractor, F^v and F^a respectively (Figure 1). The resulting features $f^v = F^v(x_i^v)$ and $f^a = F^a(x_i^a)$ are then passed to the separate classifiers G^v and G^a, whose outputs correspond to distinct score predictions (one for each modality). The final prediction results from the combination of the different modality predictions of each clip (*late fusion*), followed by the average prediction over all the clips.

4 Test-Time Adaptation for Action Recognition

In this section, we describe the proposed approach, consisting in the extension of RNA-Net to the TTA scenario (RNA^{++}) and its combination with losses aiming at reducing the classifier's uncertainty on test data (Class Relative (CR) losses).

4.1 Multi-Modal Test-Time Adaptation

A very recent work [29] showed that exploiting the multi-modal nature of videos allows one to exploit the shared knowledge available from multiple sources to build a model able to generalize to unseen data. The same strategy has also been shown to be effective as an adaptation technique when using unlabeled target data [30]. In particular, authors of [29] brought to light that the discrepancy between the two modalities' mean feature norms inhibits the network from learning equally from the two during training, i.e., the network privileges the modality with greater feature norm, while penalizing the other. This causes the final model to perform sub-optimally in comparison to the uni-modal one, a problem which has also been shown in [44]. Authors of [29] address it by proposing an audio-visual loss which minimizes the discrepancy between the two modalities' feature

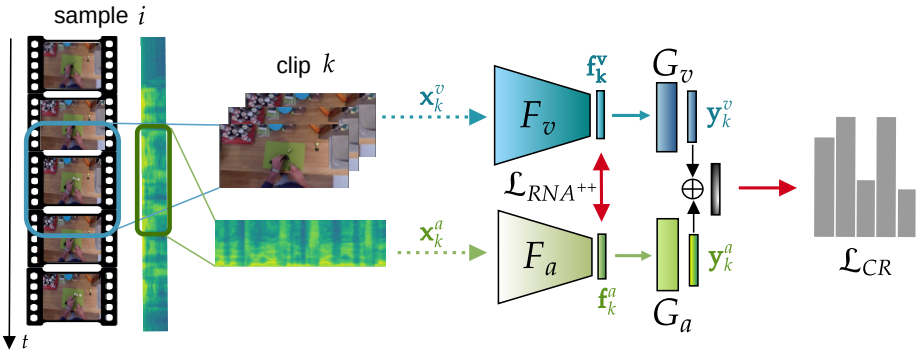

Fig. 1. Unlabeled test visual x_k^v and audio x_k^a inputs for the k-th clip are fed to the respective feature extractors F^v and F^a. The $\mathcal{L}_{RNA^{++}}$ loss operates at feature-level by balancing the relative feature norms of the two modalities. The latter is combined with Class Relative losses $\mathcal{L}_{CR}$, namely $\mathcal{L}_{MCC}$, $\mathcal{L}_{CENT}$ or $\mathcal{L}_{IM}$ losses.

norms during training, in order to better exploit multi-modal learning and thus leading to better generalization results.

This problem, referred to as the "norm-unbalance problem", still exists at test-time, negatively affecting the final prediction. In fact, the latter is biased towards the modality with greater feature norm [47], reducing the potentiality of having multiple modalities. For these reasons, in this work, we extend the proposed *Relative Norm Alignment (RNA)* loss to re-balance the mean feature norms of the two modalities during testing. This loss, which we call RNA^{++}, is designed to deal with the multi-clip nature of the test phase and it is defined as

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

where $h(X_k^m) = (\|\cdot\|_2 \circ f_k^m)(X_k^m)$ indicates the L_2-norm of the features f_k^m, $\mathbb{E}[h(X_k^m)] = \frac{1}{N} \sum_{x_{ik}^m \in \mathcal{X}_k^m} h(x_{ik}^m)$ with k the k-th clip of the m-th modality and N denotes the number of samples of the test set $\mathcal{X}_k^m = \{x_{1k}^m, ..., x_{Nk}^m\}$.

4.2 Class Relative losses

By operating at feature level, the RNA^{++} loss promotes the cooperation between the two modalities, increasing the robustness of their final embeddings and, as a result, leading to a more robust classifier which is less affected by the domain shift. However, as the RNA^{++} loss is not backpropagated through the classifier, it focuses only on the multi-modal embeddings and ignores the classification layer's uncertainty on target data. To tackle this weakness, a natural choice might be to introduce in our multi-modal framework the standard *entropy loss* [13], which is commonly used to minimize prediction *uncertainty*. However, the entropy term alone is insufficient to provide stability, as a trivial solution is the one in which the predicted single-class samples may prevail over the others [12,45], especially when dealing with unbalanced datasets. It has also been proven that the entropy

loss is not able to correctly measure the "class confusion" between correct and ambiguous classes [18]. As a result, this classifier's prediction uncertainty tends to introduce noise in the multi-clip prediction, as wrong clip predictions might dominate the correct one. Based on these considerations, minimizing the entropy is not sufficient to reduce the uncertainty of the final classifier on test samples. Thus, we re-purpose losses that bring attention to the relation between all per-class predictions in order to reduce uncertainty and we refer to them as Class Relative losses (CR losses). It follows a detailed description of these methods.

Minimum Class Confusion (MCC). This loss [18] minimizes the *inter-class confusion* on test data so that no samples are ambiguously classified into two classes at the same time. It is formalized as:

$$\mathcal{L}_{MCC} = \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} \sum_{j' \neq j}^{|\mathcal{C}|} |\widetilde{\mathbf{C}}_{jj'}| \tag{2}$$

where $\mathcal{C}$ is the number of classes and $\tilde{\mathbf{C}}_{jj'}$ measures the confusion between each class pair (j, j'). The latter is derived from the **Class Correlation** term $\mathbf{C}_{jj'}$, which is defined as:

$$\mathbf{C}_{jj'} = \hat{y}_{.j}^{T} \mathbf{W} \hat{y}_{.j'} \tag{3}$$

where we denote with $\hat{y}_{.j'}$ the j-th column of the probability matrix $\hat{Y}_{ij}$, which represents probability of the i-th samples to belong to the j-th class. $\hat{Y}_{ij}$ is obtained by summing the audio and visual probability matrices $\hat{Y}^a_{ij}$ and $\hat{Y}^v_{ij}$ respectively. The diagonal matrix $\mathbf{W}$ is used to re-weight $\mathbf{C}_{jj'}$ in order to emphasize the class with the highest class ambiguity. Finally, $\tilde{\mathbf{C}}_{jj'}$ is obtained by *category normalization* of the $\mathbf{C}_{jj'}$ value as in [41].

Information Maximization (IM). The objective of IM loss [12,34,14] is to make test-time predictions individually certain and *globally diverse* to avoid trivial solutions caused by entropy minimization alone. Indeed, it combines a conditional entropy term and a diversity term:

$$\mathcal{L}_{div} = -\mathbb{E}_{x \in \mathcal{X}} \sum_{c=1}^{\mathcal{C}} \sigma_c(h(x)) \log \sigma_c(h(x)) + \sum_{c=1}^{\mathcal{C}} \bar{p}_c \log \bar{p}_c \tag{4}$$

where $h(x) = G^v(F^v(x^v)) + G^a(F^a(x^a))$ is the C-dimensional output of each sample, summed over each modality input, and $\bar{p} = \mathbb{E}_{x \in \mathcal{X}}[\sigma(h(x))]$ is the mean of the softmax outputs for the current batch.

Complement Entropy (CENT). Considering our setting where multiple clip predictions are considered during test, the CENT loss aims at neutralizing the negative effects of incorrectly predicted clips on the final prediction. It accomplishes this by "flattening" the predicted probabilities of "complement classes", i.e., all classes except the predicted one. As a result, when several clip predictions are considered, the voting process' noise 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. Given the k-th clip, it is defined as:

$$\mathcal{L}_{CENT} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(\hat{y}_{i\bar{c}}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,j\neq p}^{\mathcal{C}}(\frac{\hat{y}_{ij}}{1-\hat{y}_{ip}}\log\frac{\hat{y}_{ij}}{1-\hat{y}_{ip}}) \quad (5)$$

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 [4], and we extend it to operate in an unsupervised fashion. In our multi-modal setting, $\hat{y}_{ij}$ results from the sum of audio and visual predictions $\hat{y}^a_{ij}$ and $\hat{y}^v_{ij}$ respectively.

5 Exprimental Results

In this section, we first introduce the dataset and the experimental setup, followed by a brief overview of the baseline methods used (Section 5.1). Finally, we present the experimental results (Section 5.2).

5.1 Experimental Setting

Dataset. We use the EPIC-Kitchens-55 dataset [8] and we adopt the same experimental protocol of [25], where the three kitchens are handpicked from the 32 available. We refer to them here as D1, D2, and D3 respectively.

Input. For RGB, during inference, 5 equidistant clips of 16 frames are fed to the network. During adaptation, we apply random crops, scale jitters and horizontal flips for data augmentation, while at pure inference time only center crops are applied. Regarding aural information, we follow [19] and convert the audio track into a 256×256 matrix representing the log-spectrogram of the signal. As for visual information, 5 equidistant audio clips in correspondence to the visual ones are used during both adaptation and inference.

Implementation Details. Our network is composed of two streams, one for each modality m, with distinct feature extractor F^m and classifier G^m. The RGB stream uses I3D [3] as in [25]. The audio feature extractor uses the BN-Inception model [15] pretrained on ImageNet, which proved to be a reliable backbone for processing audio spectrograms [19]. Each F^m produces a 1024-dimensional representation f_m which is fed to the classifier G^m, consisting in a fully-connected layer that outputs the score logits. Then, the two modalities are fused by summing the outputs. During adaptation, the network is optimized with a batch size of 32, SGD optimizer with momentum 0.1, and weight decay $1e^{-7}$. We optimized the learning rate $lr \in \{1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}\}$, loss weights $\alpha, \beta, \gamma, \delta, \epsilon$[1] $\in \{1, 0.1, 0.5, 0.01\}$, and optimization steps $n \in \{1, ..., 10\}$ for all methods, reporting the accuracy scores averaged on three different runs. The code is implemented using Pytorch framework and the models are trained using Intel(R) Core(TM) i7-9800X CPU and two GPUs Titan RTX (24 GB).

[1] $\alpha, \beta, \gamma, \delta, \epsilon$ are the weights of RNA^{++}, MCC, ENT, IM and CENT losses respectively.

Baseline Methods. We adapted the most popular image-based TTA methods to our video scenario. Those are:

- **Prediction-time BN**: authors of [48,26,33] proved that either updating [48,33] or replacing [26] batch normalization statistics μ and σ^2 with the ones from test data during inference achieves good adaptation results. In our experiments, we do not entirely replace the source statistics, but we rather update them with the ones from the target.
- **TENT** [42]: the adaptation is performed by optimizing the modulation parameters γ and β of the Batch Normalization (BN) layers by minimizing the entropy loss [13]. The normalization statistics μ and σ^2 are initialized on source data and updated for each layer in turn, during the forward pass, on test data batch statistics. We also tried a different variation of TENT, which we refer to as TENT-C, where we also optimize the classifier.
- **T3A** [16]: it is a backpropagation-free method which adjusts the classifier at test-time. In particular, it creates a pseudo-prototype for each class using online test data and the classifier pre-trained source, and then classifies each test sample basing on its distance to the pseudo-prototype.

5.2 Results

In this section, we evaluate TTA results by considering both (i) a network which has been trained on multiple source domains (*DeepAll*) and (ii) a network which has been trained with RNA-Net [29], a method which aims to improve generalization results by exploiting audio-visual correlations at feature level. On the two, we evaluate three different approaches: (i) **baseline methods**, which are standard image-based TTA methods which we adapted to our setting, (ii) **RNA^{++}**, the extension of RNA loss to operate *at feature level* on test data and (iii) **Class Relative (CR) losses**, which are losses operating *at prediction level*.

Baseline methods. We show in Table 2 and Table 3 the effects of applying existing TTA methods, namely BN [48,33], TENT [42], and T3A [16]. Despite its simplicity, BN shows a consistent improvement over both the DeepAll and RNA-Net baselines. This proves that the feature distribution varies greatly from source to target domains, an thus simply updating batch normalization statistics with the ones from target data is effective in coping with the domain shift. Both TENT and TENT-C improve over the baselines, showing that methods inherited from the image-based domain scale well to our multi-modal action recognition setting. TENT-C achieves slightly better results than TENT, proving that optimizing the classifier parameters using target data is effective in improving generalization. Both techniques benefit from having a model that has been pre-trained using a multi-modal DG strategy, since the improvement on RNA-Net is more consistent than the improvement on DeepAll.

However, the improvements are limited and slighty lower than the one obtained by BN, confirming the difficulties of this task in the egocentric context and the importance of this new benchmark to promote future reaseach in this new field. We can further notice that, differently from the others approches,

Table 2. Top-1 Accuracy (%) of different test-time adaptation methods in a Multi-Source DG scenario when applied to a **DeepAll baseline**. $D_i, D_j \rightarrow D_k$ indicates that we trained on D_i and D_j and we tested on D_k.

	D2, D3 → D1	D3, D1 → D2	D1, D2 → D3	Mean	Gain
DeepAll	51.34	43.22	41.07	45.21	-
BN [48,33]	50.82	44.14	43.91	46.29	▲+1.08
TENT [42]	49.86	42.99	43.96	45.60	▲+0.39
TENT-C [42]	49.83	43.07	44.00	45.63	▲+0.42
T3A [16]	40.28	36.86	39.24	38.79	▼−6.42
RNA^{++}	50.79	43.91	43.87	46.19	▲+0.98
ENT	51.81	43.60	43.33	46.25	▲+1.04
MCC	52.09	44.06	43.11	**46.42**	▲+**1.21**
IM	50.38	43.68	44.40	46.15	▲+0.94
CENT	51.10	43.30	44.84	46.41	▲+1.20
ENT+RNA^{++}	50.86	43.60	43.91.	46.12	▲+0.91
MCC+RNA^{++}	51.88	44.06	44.00	**46.65**	▲+**1.44**
IM+RNA^{++}	51.95	43.52	44.44	46.64	▲+1.43
CENT+RNA^{++}	50.58	43.45	45.42	46.48	▲+1.28

Table 3. Top-1 Accuracy (%) of different test-time adaptation methods in a Multi-Source DG scenario when applied to **RNA-Net baseline**.

	D2, D3 → D1	D3, D1 → D2	D1, D2 → D3	Mean	Gain
RNA-Net [29]	55.75	46.67	50.53	50.98	-
BN [48,33]	57.56	46.90	52.04	52.17	▲+1.18
TENT [42]	54.18	47.43	53.29	51.63	▲+0.65
TENT-C [42]	54.28	47.89	53.24	51.81	▲+0.83
T3A [16]	49.69	41.15	34.58	41.81	▼−9.17
RNA^{++}	57.56	46.90	52.00	52.15	▲+1.17
ENT	57.46	46.51	52.09	52.02	▲+1.04
MCC	57.70	47.05	52.04	52.26	▲+1.28
IM	57.46	47.13	52.09	52.23	▲+1.25
CENT	57.53	47.43	52.31	**52.42**	▲+**1.44**
ENT+RNA^{++}	57.29	46.67	52.04	52.00	▲+1.02
MCC+RNA^{++}	57.63	46.90	52.04	52.19	▲+1.21
IM+RNA^{++}	57.56	47.20	52.18	52.31	▲+1.33
CENT+RNA^{++}	57.67	47.51	52.18	**52.45**	▲+**1.47**

T3A does not scale to this setting. Indeed, this is explainable by the fact that the dataset used is strongly unbalanced, and a method which exploits a per-class pseudo-prototype representation could lead to sub-optimal results making prediction of classes with fewer samples almost impossible.

RNA^{++}. In Table 2 and Table 3 we illustrate the effects of minimizing RNA^{++} at test-time. It can be seen that RNA^{++} outperforms the baseline DeepAll and RNA-Net by 0.98% and 1.17% respectively, showing that re-balancing the mean feature norms of the two modalities on test samples further improves the adaptation ability of the network. It can be noticed also that starting from the robust initialization of RNA-Net to perform the rebalancing operation, it helps the RNA++ to be more effective at test time. However the limited improvement of RNA++ over existing techniques, particularly when compared to BN, highlights its need to be guided by a loss that acts on the final prediction.

CR losses. We show the performance of the entropy loss w.r.t. CR losses in Table 2 and Table 3. When applying the entropy loss, we fine-tune all the network. The entropy loss surpasses both DeepAll and RNA-Net baselines and yields results comparable to all existing TTA approaches. Except for one case, the proposed CR losses (MCC, IM, and CENT losses) surpass the entropy loss. This proves the limitation of entropy loss in this context (see Section 4.2) and highlights the benefits of using losses which take into account all the predictions.

Combining RNA^{++} with CR losses. When further combining those losses with RNA^{++}, performance increase on both settings in almost all configurations. This confirms the effectiveness of fine-tuning the network through a loss which operates not only on features but also on predictions. Indeed, the combination of RNA^{++} with CR losses proved to be the most effective technique.

The combination of RNA^{++} with the entropy loss does improves over RNA^{++} alone, while on the other side the combination of it with CR losses outperforms RNA^{++} in all cases. This provides additional evidence to support the mentioned limits of entropy (see Section 4.2) in the TTA scenario.

6 Conclusions

In this work, we investigate the test-time adaptation setting for audio-visual egocentric action recognition. We propose a new benchmark for this context, showing the performance of current image-based test-time adaptation algorithms which we adapted to the video domain. Moreover, we propose RNA^{++}, a new test-time adaptation approach which extends RNA-Net, a recent multi-modal domain generalization method. Finally, we prove the importance of combining it with a set of losses meant to further reduce classifier's uncertainty on test data. We regard our work as a starting point for future research into new settings that allow action recognition algorithms to be applied in real-world scenarios.

Acknowledgements. This work was supported by the CINI Consortium through the VIDESEC project .

References

1. Azimi, F., Palacio, S., Raue, F., Hees, J., Bertinetto, L., Dengel, A.: Self-supervised test-time adaptation on video data. In: WACV. pp. 3439–3448 (2022)
2. Bucci, S., D'Innocente, A., Liao, Y., Carlucci, F.M., Caputo, B., Tommasi, T.: Self-supervised learning across domains. TPAMI (2021)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
4. Chen, H.Y., Wang, P.H., Liu, C.H., Chang, S.C., Pan, J.Y., Chen, Y.T., Wei, W., Juan, D.C.: Complement objective training. arXiv:1903.01182 (2019)
5. Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., Zheng, J.: Temporal attentive alignment for large-scale video domain adaptation. In: CVPR. pp. 6321–6330 (2019)
6. woo Choi, J., Sharma, G., Schulter, S., Huang, J.: Shuffle and attend: Video domain adaptation. In: ECCV (2020)

7. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: Mars: Motion-augmented rgb stream for action recognition. In: CVPR (June 2019)
8. 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: ECCV. pp. 720–736 (2018)
9. Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. NIPS **32**, 6450–6461 (2019)
10. Furnari, A., Farinella, G.: Rolling-unrolling lstms for action anticipation from first-person video. TPAMI (2020)
11. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
12. Gomes, R., Krause, A., Perona, P.: Discriminative clustering by regularized information maximization. NIPS (2010)
13. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NIPS. vol. 367, pp. 281–296 (01 2004)
14. Hu, W., Miyato, T., Tokui, S., Matsumoto, E., Sugiyama, M.: Learning discrete representations via information maximizing self-augmented training. In: ICML. pp. 1558–1567 (2017)
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456 (2015)
16. Iwasawa, Y., Matsuo, Y.: Test-time classifier adjustment module for model-agnostic domain generalization. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) NIPS (2021)
17. Jamal, A., Namboodiri, V.P., Deodhare, D., Venkatesh, K.: Deep domain adaptation in action space. In: BMVC (2018)
18. Jin, Y., Wang, X., Long, M., Wang, J.: Minimum class confusion for versatile domain adaptation. In: ECCV. pp. 464–480 (2020)
19. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (October 2019)
20. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Slow-fast auditory streams for audio recognition. In: ICASSP. pp. 855–859 (2021)
21. 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: ICCV. pp. 13618–13627 (2021)
22. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive batch normalization for practical domain adaptation. Pattern Recognition **80**, 109–117 (2018)
23. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: ICCV. pp. 7083–7093 (2019)
24. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML. pp. 97–105 (2015)
25. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR (June 2020)
26. Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., Snoek, J.: Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv preprint arXiv:2006.10963 (2020)
27. Pan, B., Cao, Z., Adeli, E., Niebles, J.C.: Adversarial cross-domain action recognition with co-attention. In: AAAI. vol. 34, pp. 11815–11822 (2020)
28. Planamente, M., Bottino, A., Caputo, B.: Self-supervised joint encoding of motion and appearance for first person action recognition. In: ICPR. pp. 8751–8758 (2021)

29. Planamente, M., Plizzari, C., Alberti, E., Caputo, B.: Domain generalization through audio-visual relative norm alignment in first person action recognition. In: WACV. pp. 1807–1818 (January 2022)
30. Plizzari, C., Planamente, M., Alberti, E., Caputo, B.: Polito-iit submission to the epic-kitchens-100 unsupervised domain adaptation challenge for action recognition. arXiv preprint arXiv:2107.00337 (2021)
31. Plizzari, C., Planamente, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., Caputo, B.: E 2(go) motion: Motion augmented event stream for egocentric action recognition. arXiv preprint arXiv:2112.03596 (2021)
32. Rodin, I., Furnari, A., Mavroedis, D., Farinella, G.M.: Predicting the future from first person (egocentric) vision: A survey. CVIU p. 103252 (2021)
33. Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., Bethge, M.: Improving robustness against common corruptions by covariate shift adaptation. arXiv preprint arXiv:2006.16971 (2020)
34. Shi, Y., Sha, F.: Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. arXiv preprint arXiv:1206.6438 (2012)
35. Song, X., Zhao, S., Yang, J., Yue, H., Xu, P., Hu, R., Chai, H.: Spatio-temporal contrastive domain adaptation for action recognition. In: CVPR. pp. 9787–9795 (June 2021)
36. Sudhakaran, S., Escalera, S., Lanz, O.: Lsta: Long short-term attention for egocentric action recognition. In: CVPR. pp. 9954–9963 (2019)
37. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: ICML. pp. 9229–9248. PMLR (2020)
38. Thapar, D., Nigam, A., Arora, C.: Anonymizing egocentric videos. In: ICCV. pp. 2320–2329 (2021)
39. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
40. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. In: NIPS. pp. 5334–5344 (2018)
41. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing **17**(4), 395–416 (2007)
42. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
43. 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: ECCV. pp. 20–36 (2016)
44. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: CVPR. pp. 12695–12705 (2020)
45. Wu, X., Zhou, Q., Yang, Z., Zhao, C., Latecki, L.J., et al.: Entropy minimization vs. diversity maximization for domain adaptation. arXiv:2002.01690 (2020)
46. Yao, Z., Wang, Y., Wang, J., Yu, P., Long, M.: Videodg: Generalizing temporal relations in videos to novel domains. TPAMI (2021)
47. Ye, J., Lu, X., Lin, Z., Wang, J.Z.: Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers. ICLR (2018)
48. You, F., Li, J., Zhao, Z.: Test-time batch statistics calibration for covariate shift. arXiv preprint arXiv:2110.04065 (2021)
49. Zhao, J., Snoek, C.G.: Dance with flow: Two-in-one stream action detection. In: CVPR. pp. 9935–9944 (2019)