Toward human-robot cooperation: unsupervised domain adaptation for egocentric action recognition

Toward human-robot cooperation: unsupervised domain adaptation for egocentric action recognition

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

[1] Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129, Torino, Italy
name.surname@polito.it,
[2] Istituto Italiano di Tecnologia, Via Morego, 30, 16163, Genova, Italy,
name.surname@iit.it,
[3] Consortium Cini, Italy.

Abstract. With the advent of collaborative manipulators, the community is pushing the limits of human-robot interaction with novel control, planning, and task allocation strategies. For a purposeful interaction, however, the robot is also required to understand and predict the action of the human not only at a kinematic level (i.e. motion estimation), but also at an higher level of abstraction (i.e. action recognition), ideally from the human own perspective. Dealing with egocentric videos comes with the benefit that the data source already embeds an intrinsic attention mechanism, driven by the focus of the user. However, the deployment of such technology in realistic use-cases cannot ignore the large variability of background characteristics when changing environment, resulting in a domain shift in features space not learnable from labels at training time. In this paper, we discuss a method to perform Domain Adaptation with no external supervision, which we test on the EPIC-Kitchens-100 UDA Challenge in Action Recognition. More specifically, we move from our previous work on Relative Norm Alignment and extend the approach to unlabelled target data, enabling a simpler adaptation of the model to the target distribution in an unsupervised fashion. To this purpose, we enhanced our framework with multi-level adversarial alignment and with a set of losses aimed at reducing the classifier's uncertainty on the target data. Extensive experiments demonstrate how our approach is capable to perform Multi-Source Multi-Target Domain Adaptation, thus minimising both temporal (i.e. different recording times) and environmental (i.e. different kitchens) biases.

Keywords: Human-Robot Cooperation, First Person Action Recognition, Unsupervised Domain Adaptation

1 Introduction

Current robotics research demonstrated a significant trend in the development of technologies to support the physical interaction between humans and machines, ranging from the planning and control [2], up to their social impact [26]. However, the deployment of such technology in the real world, e.g. in household or industrial environments, requires an extension of the human intention retrieval capabilities of robots, from a mere pose estimation and forecast, to an high level description of the action executed. As an

example, considering a companion robot assisting the human in preparing a meal, the feasibility to infer from video the current action performed can enable the prediction of the next steps of the receipt, and eventually assist the cook with proper tools handover. To reach this goal, a very promising solution relies on the usage of egocentric vision, in which the human activity is recorded by wearable cameras placed on the head of the user [50]. In contrast with standard third person Computer Vision (CV) tasks, this setting comes with the benefit that source data are characterised by a rich multi-modal information, thanks to the proximity of audio/video sensors to the action scene, and by an intrinsic embedding of an attention mechanisms that stems from the human gaze direction itself. Although egocentric vision rapidly attracted the interest of the research community [60,58,21,30,23,66], this particular setup of data collection comes with several difficulties: i) ego-motions represents a significant source of noise for the dataset, because changes in head posture cause a shift in the point-of-view and background, introducing confusion between ego-motion and the real action of the subject; 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*)[42], thus resulting in decreased performances when the environment changes (e.g. different kitchens); iii) video recordings of actions can change in time, e.g. as a consequence to differences in illumination (night vs. day), habits, or changes in human skills on the longer distance.

It is important to recall that the effect of this problem is not consistent across different sensing modalities. Considering, as an example, the ego-motion, the impact on the auditory channel is extremely limited, while the visual domain is strongly affected. The optical flow, instead, is more focused on the motion in the scene, rather than the appearance, and is therefore less sensitive to environmental changes [42] (see e.g. Fig. 1). Despite being more subject to environmental bias, RGB data are richer information sources, representing in detail all the objects present in the scene. As such, they play a crucial role for understanding the affordances of the scene [24]. Lastly, the domain shift of the audio signal is further distinct from the visual one (e.g., the sound of 'cut' will differ from a plastic to a wooden cutting board). These observations suggest that domain shifts are not all of the same nature, and their impact can vary significantly across modalities. As a consequence, it is of crucial relevance to develop classifiers able to assess - depending on the conditions - which modality is more informative, and therefore modulate on the flight the weights that combine different sensing inputs for the output definition. This approach demonstrated how increasing the network's multi-modal learning capability promotes model resilience across domains, allowing the model to better recognize action under diverse domain shifts.

As a step in this direction, in our previous work [46], we 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, our results demonstrated that a mere feeding of 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[46]. However, it must be noted that the RNA-Net

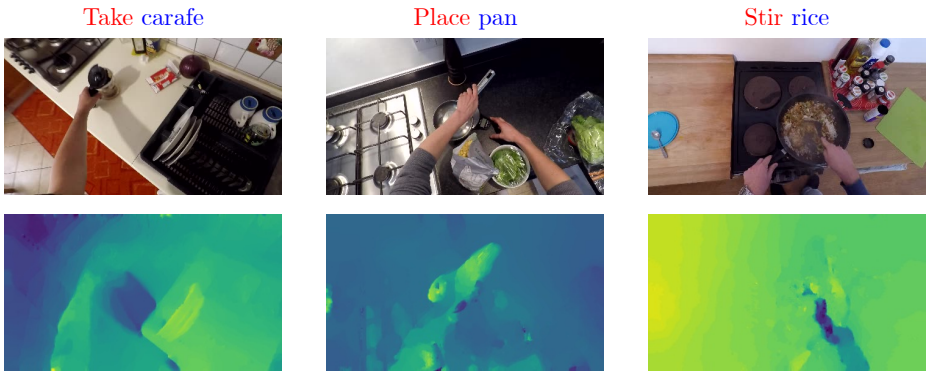

Fig. 1. Three examples of RGB frames (top line) and colorized optical flow (bottom line) with the corresponding verb and noun labels. The three frames come from different kitchens in the Epic-Kitchens-100 dataset.

assumes a simplified (and unrealistic) scenario, in which only one single domain shift (environmental bias, i.e. different kitchens) and few different actions (8 labels) are considered. With this work, we extend and generalise the field of application of RNA-Net to a more realistic and challenging scenario, where i) most of the possible domain shifts are considered and ii) we considerably increase the number of actions and objects available for the classifier (up to 97 actions and 300 different objects).

To reach this goal, in this work we developed a method to tackle the most realistic setting possible, adopting the Epic-Kitchens-100 dataset and facing the corresponding UDA challenge. Indeed, both the presence of different environments and the fact that the source and target domains are captured in different temporal moments make it a multi-shift problem ideal for our experiments. Our basis idea is that, to increase the consistency of the predictions and close the source-to-target accuracy gap, it is necessary to tackle concurrently both the temporal shift and the environmental bias. Such a problem is an extension of a simple UDA setting and may be referred to as Multi-Source Multi-Target Domain Adaptation. The term ”Multi” refers to the various environments found in both the source and target datasets.

As anticipated before, to face this new setting we exploited the capability of the RNA-Net method of close the gap between feature norms of different modalities. Our method was complemented with several domain adaptation techniques which tackle other aspects of the domain shift. In particular, we integrated several branches at distinct levels of abstraction, namely frame-level and video-level, into our framework to use the adversarial alignment technique for feature adaptation as in [9]. All of the techniques presented attempt to use the target data by acting simply on the model's features; differently, we introduce in this framework a set of losses aimed at minimising the classifier's uncertainty on the target data, thereby boosting the model's adaptability. Lastly, to deal with the significant challenges posed by this dataset, we adopted different models as an ensemble to obtain the final prediction and we introduce also a set of ensemble UDA losses.

To summarize, this paper advances the state of the art with the development of a strong UDA pipeline that we tested on a competitive international challenge (Epic-Kitchens-100)[4]. Our contributions are (see also Figure 2):

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 results were obtained by combining different model streams by means of DA-based losses, namely Min-Entropy Consistency (MEC) and Complement Entropy (CENT).

2 Related Works

First Person Action Recognition. So far, most of the research effort has been focused on data provided by a specific view of the camera (often fixed), i.e., third person view [54,64,5]. With the recent release of a large-scale dataset of first-person actions [14], the community started to investigate the potential and the challenges of videos recorded from an egocentric viewpoint. As we anticipated in the Introduction, first person action recognition suffers from sudden changes of view caused by the motion of the camera. To tackle this problem, the main approaches proposed so far are based on multi-stream architectures [5,54,41,37,6,30,40], many of which are inherited from the third-person action recognition literature. The networks used to extract spatial-temporal information from egocentric videos can be divided into two main groups. The first exploits Long Short-Term Memory and variants [59,60,58,21] to generate an embedding representation based on the temporal relations between the features frames. The second [55,62,66,29] leverages 3D convolutional kernels, which jointly generate spatial-temporal features by sliding along the spatial and temporal dimensions. Recent works also exploit an attention mechanism at frame or clip level [60,58,44,39,40] to re-weight the spatial or temporal features, obtaining interesting results. By observing the importance of multi-stream approaches in this context, several papers investigate alternative methods to fuse streams w.r.t. the standard late fusion approach, creating a more compact multi-modal representation [57,65,66,69]. The most popular technique in this context is the multi-modal approach [5,42,64,58,21], especially in EPIC-Kitchens competitions [15,14]. Indeed, RGB data is frequently combined with motion data, such as optical flow and audio information [31].

Unsupervised Domain Adaptation (UDA). The goal of UDA is to bridge the domain gap between a labeled source domain and an unlabeled target one, which often are drawn from different data distributions. We can divide unsupervised domain adaptation approaches into *discrepancy-based* methods, which explicitly minimise a distance metric between source and target distributions [67,53,38], and *adversarial-based*

[4] https://epic-kitchens.github.io/2022

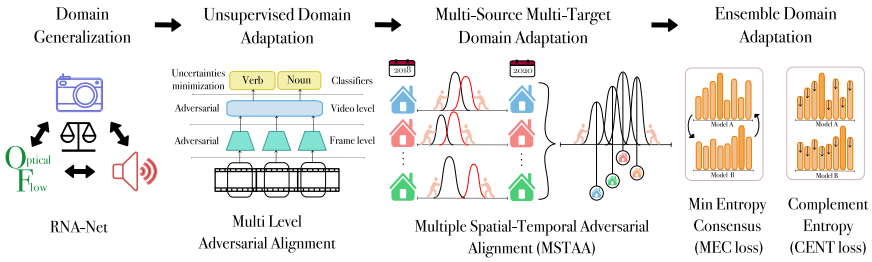

Fig. 2. An overview of the proposed approach. It can be summarized in four main aspects: **1.** Domain Generalization through RNA-Net [46], **2.** Unsupervised Domain Adaptation via Multi-Level Adversarial Alignment and entropy minimization, **3.** Multi-Source Multi-Target Domain Adaptation extension and **4.** Ensemble Domain Adaptation losses.

methods [19,61], which target the same goal using a gradient reversal layer (GRL) [22]. Other works, instead, exploit batch normalisation layers to normalise source and target statistics [35,36,7]. The approaches described above have been designed for standard image classification tasks. Considering methods developed specifically for video tasks, instead, it is worth mentioning UDA for action detection [1], segmentation [10] and classification [9,42,12,28,43,56]. To align the temporal dynamics of feature space for videos, several works on video domain adaptation use an adversarial learning framework like DAAA [28], also at multi-level such as in TA3N [9], in conjunction with an attention mechanism. TCoN [43] exploits a cross-domain co-attention mechanism to match the feature distributions between source and target domains, for temporal alignment. Other methods, instead, exploit jointly with the adversarial approach also auxiliary self-supervised tasks. As an example, in Munro et al. [42] the authors propose a synchronisation task to learn the multi-modal correspondence of RGB and optical flow. SAVA [13], instead, proposes a self-supervised predictive method for video domain adaptation, which aims to predict the clip order. Instead in [56,32,52], the authors propose a self-supervised contrastive learning approach for video domain adaptation.

Domain Generalization (DG). The DG scenario is a particular setting in which no target data is available at all, and the model is expected to learn to generalise using inputs from a single or multiple source domains, as it may happen in realistic scenarios. Previous approaches in DG are mostly designed for image data [4,63,33,20,34,3] and are divided in *feature-based* and *data-based* methods. The former focus on extracting invariant information which are shared across-domains [33,34], while the latter exploit data-augmentation strategies to augment source data with adversarial samples to get closer to the target [63]. Interestingly, using a self-supervised pretext task is an efficient solution for the extraction of a more robust data representation [4,3]. Recently, in [46] we proposed a feature-level solution for Domain Generalization problem in first person action recognition by leveraging audio-visual correlations.

3 Problem Definition

Epic-Kitchens Action Recognition Challenge, is based on a dataset of video consisting of trimmed actions, where the start and end times of each action are given.

The objective of the challenge is to understand the activity executed in each sample, uniquely defined by the couple 'verb' (i.e. the actual action) and the main interacting object ('noun'). In the official dashboard [5] the authors report the performance linked to all those 3 categories separately. This is done because objects and actions represent the two main pillars that describe the high-level activity implemented, but since they are encoded differently in the input data (mainly RGB for objecs, motion for actions), in some cases the model could be more accurate in classifying only one of the two. The overall activity classification is assumed to be correct only if both the object and the action are classified properly.

Given $ks \geq 1$ source domains $\{\mathcal{S}_1, ..., \mathcal{S}_{ks}\}$, where each $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ is composed of N_s source samples with label space Y^s known, our goal is to learn a model representation able to perform well on $kt \geq 1$ target domains $\{\mathcal{T}_1, ..., \mathcal{T}_{kt}\}$, where each $\mathcal{T} = \{x_i^t\}_{i=1}^{N_t}$ of N_t target samples whose label space Y^t is unknown. If we consider $\mathcal{D}_{s,i}, \mathcal{D}_{s,j}$ distributions of the i-th and the j-th source domain and $\mathcal{D}_{t,w}, \mathcal{D}_{t,z}$ distributions of the w-th and the z-th target domain. Our two main assumptions are that the distributions of all the domains are different, i.e., $\mathcal{D}_{s,i} \neq \mathcal{D}_{t,w} \wedge \mathcal{D}_{s,i} \neq \mathcal{D}_{s,j} \wedge \mathcal{D}_{t,w} \neq \mathcal{D}_{t,z}$, with $i \neq j$ and with $w \neq z$, $i, j = 1, ..., ks$ and $w, z = 1, ..., kt$, and that the union of all label spaces is shared, $\mathcal{C}_s = \mathcal{C}_{s,1} \cup ... \cup \mathcal{C}_{s,ks} = \mathcal{C}_{t,1} \cup ... \cup \mathcal{C}_{t,kt} = \mathcal{C}_t$. In this work we consider three different scenarios:

Domain Generalization (DG), where at training time the model can access one or more fully labeled source datasets $\mathcal{S}_1, ..., \mathcal{S}_{ks}$, but no information is available about the target domains $\mathcal{T}_1, ..., \mathcal{T}_{kt}$. The objective is to train a model able to predict an action of the target domain without having access to target data at training time, thus exploiting the knowledge from multiple source domains to improve generalization. The literature refers to this setting considering the number of target domains kt equal to 1.

Unsupervised Domain Adaptation (UDA), where at training time it is possible to access a set of unlabeled target samples belonging to the target domains $\mathcal{T}_1, ..., \mathcal{T}_{kt}$, jointly with one fully labeled source domain $\mathcal{S}_1, ..., \mathcal{S}_{ks}$. Usually the literature refers to this setting considering both the number of target domains kt and the number of source domains ks equal to 1.

Multi-Sources Multi-Target Unsupervised Domain Adaptation, an extension of the previous setting, with the only difference that ks and kt are larger than one.

4 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.

4.1 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

[5] https://codalab.lisn.upsaclay.fr/competitions/1241#results

recently proposed to operate in this context, called Relative Norm Alignment (RNA) [46]. This method 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 [46] 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 this work, 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.

4.2 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 unsupervised domain adaptation techniques for videos, we integrate into our framework an adversarial approach [9,42], consisting of an extension of the DANN [22] typical 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 [9] as in [48]. This action minimizes the entropy, resulting in a refinement of the classifier adaptation. The term "attentive" refers to a loss re-weighting approach that prioritizes videos with low domain discrepancy by focusing on minimizing entropy for these videos.

4.3 Multi-Source Multi-Target Domain Adaptation

The previous Epic Kitchen challenges [17,16], as well as the literature on unsupervised domain adaptation for first person action recognition [42,46,45,49,47], 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 [42], 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.

4.4 Ensemble UDA losses

For our final testing, 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, repurposing 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 [51] to encourage coherent predictions between different models. The resulting loss is defined as:

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

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:

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

where N is the total number of samples in the batch, $\hat{y}_{ip} = max_j(\hat{y}_{ij})$ represents the predicted probability of the class p with the higher score for the i-th sample, 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 [8], and we extend it to operate in an unsupervised fashion.

5 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.

Backbone. For our submission, we adopted three different network configurations. In the first one, corresponding to the RNA-Net framework in [46], we used the Inflated 3D ConvNet (I3D), pre-trained on Kinetics [5], for RGB and Flow streams, and a BN-Inception model [27] pre-trained on ImageNet [18] 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 Temporal-Binding-Network (TBN) [42] model trained on EPIC-Kitchens-55. In the last configurations, we used standard ResNet-50 architectures [25] equipped with the Temporal Shift Module (TSM) [37] pre-trained on EPIC-Kitchens-55 [6].

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 sub-module (SMR), proposed in [68], to generate a common frame-embedding among the modalities. Then, using temporal pooling, we obtain a final video-embedding that is sent to the verb and noun classifiers.

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

6 Experimental Setting

Dataset. EPIC-KITCHENS-100 [15] is the dataset utilized in our experiments; it was recorded by 16 individuals from diverse nations (in North America and Europe). The dataset is divided into two major groups: *Source* and *Target* data, which include 16.115 and 32.024 samples, respectively. The source data consists of labelled videos recorded in 2018, whereas the target data consists of unlabeled videos collected in 2020. Both categories are further subdivided into train, validation, and test sets. The dataset includes a total of 3369 possible actions, each of which corresponds to a combination of "verb" and "noun". The total classes for verb and noun are respectively 97 and 300.
Input. For RNA-Net, we use 16 continuous frames (segment) randomly sampled for each modality during training, while at test time 5 equidistant segments spanning across all clips are fed to the network. At training time, we apply random crops, scale jitters and horizontal flips for data augmentation, while at test time only center crops are applied. Regarding aural information, we follow [30] and convert the audio track into a 256 $\times$ 256 matrix representing the log-spectrogram of the signal. The audio clip is first extracted from the video, sampled at 24kHz and then the Short-Time Fourier Transform (STFT) is calculated of a window length of 10ms, hop size of 5ms and 256 frequency bands. Hence, the x and y axis represent time and frequency, respectively. As regards the other two architectures, TSM and TBN, we use respectively 8 and 25 frames uniformly samples along all the videos.
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 [46]. Instead, when using pre-extracted features from ResNet50 or BNInception, 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. We weighted RNA, CENT and MEC losses $\lambda_{RNA} = 1$, $\lambda_{CENT} = 0.31$ and $\lambda_{MEC} = 0.22$ respectively. In addition, we report the values used to weight the attentive entropy loss, $\gamma = 0.003$, and the domain losses at different levels for MSTAA, $\beta = (0.75, 0.75, 0.75)$.

7 Results

In Table 1 we report our best performing model on the target test, achieving the **2nd** position on 'verb', and the **3rd** on 'noun' and 'action'. The **3rd** place on 'action' obtained in the challenge is a demonstration of the robustness of the pipeline developed. Indeed, unlike the first two positions [11], our approach focuses on developing an unsupervised domain adaptation strategy that is independent of the backbone used. Indeed, the disparity between our results and the other two approaches is justified by the fact that they either used large models for action recognition (**2nd** place) or introduced hand detection as a secondary branch (**1st** place [11]). Our results, instead, demonstrate that our UDA pipeline is competitive even without adopting state-of-the-art models or auxiliary tasks (such as hand detection). Adding the techniques reported above is complementary to our work and is another step toward solving the multi-source multi-target domain adaptation realistic setting. Additionally, in Tables 2 (left and right) we show an ablation of the proposed UDA and DG methods described in section 4.

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 [15]	8	46.91	27.69	18.95	72.70	50.72	30.53
EPIC_TA3N_SOURCE_ONLY [15]	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**

Domain Generalization			
	Target	Verb Top-1	Verb Top-5
Source Only	✗	44.39	69.69
EPIC_TA3N [15]	✓	46.91	72.70
RNA-Net [46]	✗	47.96	79.54
EPIC_TA3N+RNA-Net	✓	**50.40**	**80.47**

Table 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.

How well do DG approaches perform? The results in Table 2(right) 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 [9] 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(left) 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).

8 Conclusions

In this paper, we introduced and discussed the potentiality of egocentric vision for human-robot cooperation. Indeed, this source of information may come with the interesting benefit of an intrinsic attention mechanism associated to the user head posture and gaze direction. The wearability of the sensing setup makes this technology particularly suitable to be deployed in unstructured scenarios, where robots and humans are co-existing in a daily-life environment. While from one side this is a key enabling factor for the actual deployment of human-robot cooperating frameworks, it also brings several challenges, such as the severe noise superimposition caused by ego-motion, and the strong domain shifts associated to the particular unstructured tasks (i.e. changing environments, subjects, habits). For this reason, to really unlock the potential of egocentric vision, it is important to develop models able to generalize efficiently across domains. In this paper, we presented an unsupervised approach for Multi-Source Multi-Target Domain Adaptation for egocentric action recognition as a strong solution to the problem of temporal and spatial biases. Moving from these promising results, we plan to integrate our framework in a more general human-robot cooperation framework, in which the manipulator will be able to identify the action performed by the human, and eventually plan a consequent action to support human tasks.

Acknowledgements. This work was supported both 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. Agarwal, N., Chen, Y.T., Dariush, B., Yang, M.H.: Unsupervised domain adaptation for spatio-temporal action localization. arXiv preprint arXiv:2010.09211 (2020)
2. Ajoudani, A., Zanchettin, A.M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., Khatib, O.: Progress and prospects of the human–robot collaboration. Autonomous Robots **42**(5), 957–975 (2018)
3. Bucci, S., D'Innocente, A., Liao, Y., Carlucci, F.M., Caputo, B., Tommasi, T.: Self-supervised learning across domains (2020)
4. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR, pp. 2229–2238 (2019)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR, pp. 6299–6308 (2017)
6. Cartas, A., Luque, J., Radeva, P., Segura, C., Dimiccoli, M.: Seeing and hearing egocentric actions: How much can we learn? In: ICCV Workshops, pp. 0–0 (2019)
7. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: CVPR, pp. 7354–7362 (2019)
8. 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 preprint arXiv:1903.01182 (2019)
9. Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., Zheng, J.: Temporal attentive alignment for large-scale video domain adaptation. In: ICCV, pp. 6321–6330 (2019)
10. Chen, M.H., Li, B., Bao, Y., AlRegib, G., Kira, Z.: Action segmentation with joint self-supervised temporal domain adaptation. In: CVPR, pp. 9454–9463 (2020)

11. Cheng, Y., Fang, F., Sun, Y.: Team vi-i2r technical report on epic-kitchens-100 unsupervised domain adaptation challenge for action recognition 2021. arXiv preprint arXiv:2206.02573 (2022)
12. Choi, J., Sharma, G., Chandraker, M., Huang, J.B.: Unsupervised and semi-supervised domain adaptation for action recognition from drones. In: WACV, pp. 1717–1726 (2020)
13. Choi, J., Sharma, G., Schulter, S., Huang, J.B.: Shuffle and attend: Video domain adaptation. In: ECCV, pp. 678–695. Springer (2020)
14. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset (2018)
15. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision. arXiv preprint arXiv:2006.13256 (2020)
16. Damen, D., Kazakos, E., Price, W., Ma, J., Doughty, H.: Epic-kitchens-55 - 2020 challenges report (2020)
17. Damen, D., Price, W., Kazakos, E., Furnari, A., Farinella, G.M.: Epic-kitchens - 2019 challenges report (2019)
18. Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei: Imagenet: A large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009). DOI 10.1109/CVPR.2009.5206848
19. Deng, Z., Luo, Y., Zhu, J.: Cluster alignment with a teacher for unsupervised domain adaptation. In: ICCV, pp. 9944–9953 (2019)
20. Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. NeurIPS **32**, 6450–6461 (2019)
21. Furnari, A., Farinella, G.: Rolling-unrolling lstms for action anticipation from first-person video. T-PAMI (2020)
22. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML, pp. 1180–1189. PMLR (2015)
23. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: CVPR, pp. 12,046–12,055 (2019)
24. Gibson, J.J.: The theory of affordances. Hilldale, USA **1**(2), 67–82 (1977)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
26. Henschel, A., Hortensius, R., Cross, E.S.: Social cognition in the age of human–robot interaction. Trends in Neurosciences **43**(6), 373–384 (2020)
27. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: F. Bach, D. Blei (eds.) ICML, *Proceedings of Machine Learning Research*, vol. 37, pp. 448–456. PMLR (2015). URL http://proceedings.mlr.press/v37/ioffe15.html
28. Jamal, A., Namboodiri, V.P., Deodhare, D., Venkatesh, K.: Deep domain adaptation in action space. In: BMVC, vol. 2, p. 5 (2018)
29. Kapidis, G., Poppe, R., van Dam, E., Noldus, L., Veltkamp, R.: Multitask learning to improve egocentric action recognition. In: ICCV Workshops, pp. 0–0 (2019)
30. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (2019)
31. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Slow-fast auditory streams for audio recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 855–859. IEEE (2021)
32. 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. 13,618–13,627 (2021)

33. Li, H., Jialin Pan, S., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR, pp. 5400–5409 (2018)
34. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: ECCV, pp. 624–639 (2018)
35. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive batch normalization for practical domain adaptation. Pattern Recognition **80**, 109–117 (2018)
36. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. In: ICLR. OpenReview.net (2017). URL https://openreview.net/forum?id=Hk6dkJQFx
37. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: ICCV, pp. 7083–7093 (2019)
38. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML, pp. 97–105. PMLR (2015)
39. Lu, M., Li, Z., Wang, Y., Pan, G.: Deep attention network for egocentric action recognition. IEEE Transactions on Image Processing **28**(8), 3703–3713 (2019)
40. Lu, M., Liao, D., Li, Z.N.: Learning spatiotemporal attention for egocentric action recognition. In: ICCV Workshops, pp. 0–0 (2019)
41. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: CVPR, pp. 1894–1903 (2016)
42. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR, pp. 122–132 (2020)
43. Pan, B., Cao, Z., Adeli, E., Niebles, J.C.: Adversarial cross-domain action recognition with co-attention. In: AAAI, pp. 11,815–11,822 (2020)
44. Perez-Rua, J.M., Martinez, B., Zhu, X., Toisoul, A., Escorcia, V., Xiang, T.: Knowing what, where and when to look: Efficient video action modeling with attention (2020)
45. Planamente, M., Bottino, A., Caputo, B.: Self-supervised joint encoding of motion and appearance for first person action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 8751–8758. IEEE (2021)
46. 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 (2022)
47. Plananamente, M., Plizzari, C., Caputo, B.: Test-time adaptation for egocentric action recognition. In: International Conference on Image Analysis and Processing, pp. 206–218. Springer (2022)
48. 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)
49. Plizzari, C., Planamente, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., Caputo, B.: Eˆ 2 (go) motion: Motion augmented event stream for egocentric action recognition. arXiv preprint arXiv:2112.03596 (2021)
50. Rodin, I., Furnari, A., Mavroeidis, D., Farinella, G.M.: Predicting the future from first person (egocentric) vision: A survey. Computer Vision and Image Understanding **211**, 103,252 (2021)
51. Roy, S., Siarohin, A., Sangineto, E., Bulo, S.R., Sebe, N., Ricci, E.: Unsupervised domain adaptation using feature-whitening and consensus loss. In: CVPR, pp. 9471–9480 (2019)
52. Sahoo, A., Shah, R., Panda, R., Saenko, K., Das, A.: Contrast and mix: Temporal contrastive video domain adaptation with background mixing. NeurIPS **34** (2021)
53. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR, pp. 3723–3732 (2018)
54. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS, NIPS'14, p. 568–576. MIT Press, Cambridge, MA, USA (2014)

55. Singh, S., Arora, C., Jawahar, C.: First person action recognition using deep learned descriptors. In: CVPR, pp. 2620–2628 (2016)
56. 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 (2021)
57. Sudhakaran, S., Escalera, S., Lanz, O.: Hierarchical feature aggregation networks for video action recognition. arXiv preprint arXiv:1905.12462 (2019)
58. Sudhakaran, S., Escalera, S., Lanz, O.: Lsta: Long short-term attention for egocentric action recognition. In: CVPR, pp. 9954–9963 (2019)
59. Sudhakaran, S., Lanz, O.: Convolutional long short-term memory networks for recognizing first person interactions. In: ICCV Workshops (2017)
60. Sudhakaran, S., Lanz, O.: Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. arXiv preprint arXiv:1807.11794 (2018)
61. Tang, H., Jia, K.: Discriminative adversarial domain adaptation. In: AAAI, pp. 5940–5947 (2020)
62. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV, pp. 4489–4497 (2015)
63. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. In: NeurIPS, pp. 5334–5344 (2018)
64. 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. Springer (2016)
65. Wang, X., Wu, Y., Zhu, L., Yang, Y., Zhuang, Y.: Symbiotic attention: Uts-baidu submission to the epic-kitchens 2020 action recognition challenge
66. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: CVPR (2019)
67. Xu, R., Li, G., Yang, J., Lin, L.: Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In: ICCV, pp. 1426–1435 (2019)
68. Yang, L., Huang, Y., Sugano, Y., Sato, Y.: Epic-kitchens-100 unsupervised domain adaptation challenge for action recognition 2021: Team m3em technical report. arXiv preprint arXiv:2106.10026 (2021)
69. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV, pp. 803–818 (2018)