Domain generalization through audio-visual relative norm alignment in first person action recognition

Domain Generalization through Audio-Visual Relative Norm Alignment in First Person Action Recognition

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

[1] Politecnico di Torino [2] Istituto Italiano di Tecnologia [3] CINI Consortium

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

Abstract

First person action recognition is becoming an increasingly researched area thanks to the rising popularity of wearable cameras. This is bringing to light cross-domain issues that are yet to be addressed in this context. Indeed, the information extracted from learned representations suffers from an intrinsic "environmental bias". This strongly affects the ability to generalize to unseen scenarios, limiting the application of current methods to real settings where labeled data are not available during training. In this work, we introduce the first domain generalization approach for egocentric activity recognition, by proposing a new audio-visual loss, called Relative Norm Alignment loss. It re-balances the contributions from the two modalities during training, over different domains, by aligning their feature norm representations. Our approach leads to strong results in domain generalization on both EPIC-Kitchens-55 and EPIC-Kitchens-100, as demonstrated by extensive experiments, and can be extended to work also on domain adaptation settings with competitive results.

1. Introduction

First Person Action Recognition is rapidly attracting the interest of the research community [72, 69, 27, 36, 30, 86, 25, 63, 46], both for the significant challenges it presents and for its central role in real-world egocentric vision applications, from wearable sport cameras to human-robot interaction or human assistance. The recent release of the EPIC-Kitchens large-scale dataset [18], as well as the competitions that accompanied it, has encouraged research into more efficient architectures. In egocentric vision, the recording equipment is worn by the observer and it moves around with her. Hence, there is a far higher degree of changes in illumination, viewpoint and environment compared to a fixed third person camera. Despite the numerous publications in the field, egocentric action recognition still has one major flaw that remains unsolved, known as "environmental bias" [77]. 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. To give an intuition of its impact, we show in Figure 2 the relative drop in model performance from the seen to unseen test set of the top-3 methods of the 2019 and 2020 EPIC-Kitchens challenges. In general, this problem is referred to in the literature as *domain shift*, meaning that a model trained on a source labelled dataset cannot generalize well on an unseen dataset, called target. Recently, [56] addressed this issue by reducing the problem to an unsupervised domain adaptation (UDA) setting, where an unlabeled set of samples from the target is available during training. However, the UDA scenario is not always realistic, because the target domain might not be known a priori or because accessing target data at training time might be costly (or plainly impossible).

In this paper, we argue that the true challenge is to learn a representation able to generalize to any unseen domain, regardless of the possibility to access target data at training time. This is known as domain generalization (DG) setting.

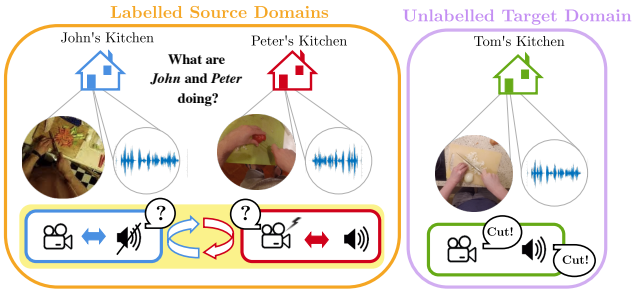

Figure 1. Due to their intrinsic different nature, audio and visual information suffer from the domain shift in different ways. However, the network tends to "privilege" one modality over the other. We re-balance the contribution of the two modalities while training, allowing the network to learn "equally" from each.

*The authors equally contributed to this work.

Inspired by the idea of exploiting the multi-modal nature of videos [56, 36], we make use of multi-sensory information to deal with the challenging nature of the setting. Although the optical flow modality is the most widely utilized [56, 82, 69, 27], it requires a high computational cost, limiting its use in online applications. Furthermore, it may not be ideal in a wearable context where battery and processing power are restricted and must be preserved. The audio signal has the compelling advantage of being natively provided by most wearable devices [43], and thus it does not require any extra processing. Egocentric videos come with rich sound information, due to the strong hand-object interactions and the closeness of the sensors to the sound source, and audio is thus suitable for first person action recognition [36, 10, 37]. Moreover, the "environmental bias" impacts auditory information as well, but in a different way than it affects visual information. In fact, audio and video originate from distinct sources, i.e., camera and microphone. We believe that the complementarity of the two can help to attenuate the domain shift they both suffer. For instance, the action "cut" presents several audio-visual differences across domains: cutting boards will differ in their visual and auditory imprints (i.e., wooden cutting board *vs* plastic one), various kinds of food items might be cut, and so forth (Figure 1).

Despite multiple modalities could potentially provide additional information, the CNNs' capability to effectively extract useful knowledge from them is somehow restricted [83, 3, 31, 61, 85]. The origin of this difficulty, in our opinion, is due to one modality being "privileged" over the other during training. Motivated by these findings, we propose the Relative Norm Alignment loss, a simple yet effective loss whose goal is to re-balance the mean feature norms of the two modalities during training, allowing the network to fully exploit joint training, especially in cross-domain scenarios. To summarize, our contributions are the following:

- we bring to light the "unbalance" problem arising from training multi-modal networks, which causes the network to "privilege" one modality over the other during training, limiting its generalization ability;
- we propose a new cross-modal audio-visual loss, the Relative Norm Alignment (RNA) loss, that progressively aligns the relative feature norms of the two modalities from various source data, resulting in domain-invariant audio-visual features;
- we present a new benchmark for multi-source domain generalization in first person videos and extensively validate our method on both DG and UDA scenarios.

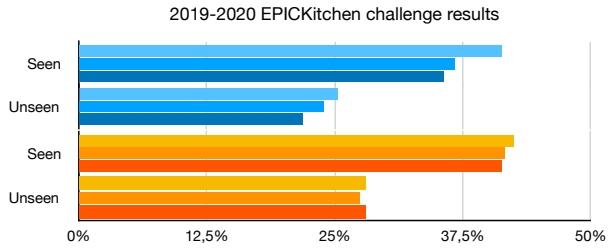

Figure 2. Top-3 results of the 2019 [21] and 2020 [20] EK challenges, when testing on "Seen" and "Unseen" kitchens.

2. Related Works

First Person Action Recognition. The main architectures utilized in this context, which are generally inherited from third-person literature, divide into two categories: methods based on 2D convolution [66, 82, 52, 50, 93, 36, 10, 70] and method based on 3D ones [9, 67, 86, 35, 78, 26, 66, 56]. LSTM or its variations [71, 72, 69, 27, 60] commonly followed the first group to better encode temporal information. The most popular technique is the multi-modal approach [9, 56, 82, 69, 27], especially in EPIC-Kitchens competitions [19, 18]. Indeed, RGB data is frequently combined with motion data, such as optical flow. However, although optical flow has proven to be a strong asset for the action recognition task, it is computationally expensive. As shown in [17], the use of optical flow limits the application of several methods in online scenarios, pushing the community either towards single-stream architectures [91, 17, 44, 73, 60], or to investigate alternative modalities, e.g., audio information [37]. Although the audio modality has been proven to be robust in egocentric scenarios by [36, 10, 37], this work is the first to exploit it, jointly with its visual counterpart, in a cross-domain context.

Audio-Visual Learning. Many representation learning methods use self-supervised approaches to learn cross-modal representations that can be transferred well to a series of downstream tasks. Standard tasks are Audio-Visual Correspondence [5, 41, 6], or Audio-Visual Synchronization [16, 1, 58, 42], which was shown to be useful for sound-source localization [6, 1, 90, 65, 76], active speaker detection [16, 1] and multi-speaker source separation [58, 1]. Other audio-visual approaches have been recently proposed [29, 53, 75, 54, 55, 4, 40] which exploit the natural correlation between audio and visual signals. Audio has also attracted attention in egocentric action recognition [36, 10] and has been used in combination with other modalities [83]. However, none of these techniques has been intended to cope with cross-domain scenarios, whereas this paper demonstrates the audio modality's ability to generalize to unseen domains when combined with RGB.

Unsupervised Domain Adaptation (UDA). We can divide UDA approaches into *discrepancy-based* methods,

which explicitly minimize a distance metric among source and target distributions [87, 64, 51], and *adversarial-based* methods [23, 74], often leveraging a gradient reversal layer (GRL) [28]. Other works exploit batch normalization layers to normalize source and target statistics [48, 49, 11]. The approaches described above have been designed for standard image classification tasks. Other works analyzed UDA for video tasks, including action detection [2], segmentation [13] and classification [12, 56, 15, 34, 59, 68]. Recently [56] proposed an UDA method for first person action recognition, called MM-SADA, consisting of a combination of existing DA methods trained in a multi-stage fashion.

Domain Generalization (DG). Previous approaches in DG are mostly designed for image data [8, 79, 45, 24, 47, 7] and are divided in *feature-based* and *data-based* methods. The former focus on extracting invariant information which are shared across-domains [45, 47], while the latter exploit data-augmentation strategies to augment source data with adversarial samples to get closer to the target [79]. Interestingly, using a self-supervised pretext task is an efficient solution for the extraction of a more robust data representation [8, 7]. We are not aware of previous works on first or third person DG. Among unpublished works, we found only one *arXiv* paper [88] in third person action recognition, designed for single modality. Under this setting, first person action recognition models, and action recognition networks in general, degenerate in performance due to the strong divergence between source and target distributions.

Our work stands in this DG framework, and proposes a feature-level solution to this problem in first person action recognition by leveraging audio-visual correlations.

3. Proposed Method

In this work, we bring to light that the discrepancy between the ***mean feature norms*** of audio and visual modalities negatively affects the training process of multi-modal networks, leading to sub-optimal performance. Indeed, it causes the modality with greater feature norm to be "privileged" by the network during training, while "penalizing" the other. We refer to this problem with the term ***"norm unbalance"***. The intuitions and motivations behind this problem, as well as our proposed solution to address it, are presented below.

3.1. Intuition and Motivation

A common strategy in the literature to solve the first-person action recognition task is to use a multi-modal approach [56, 36, 50, 10, 37, 82]. Despite the wealth of information of multi-modal networks w.r.t. the uni-modal ones, their performance gains are limited and not always guaranteed [83, 3, 31, 61, 85]. Authors of [83] attributed this problem to overfitting, and addressed it by re-weighting the loss value of each stream through different hyperparameters. This technique, however, necessitates a precise estimating step, which is dependent on the task and the dataset. In this paper, we approach the above described multi-modal issue from a different perspective, by considering an audio-visual framework.

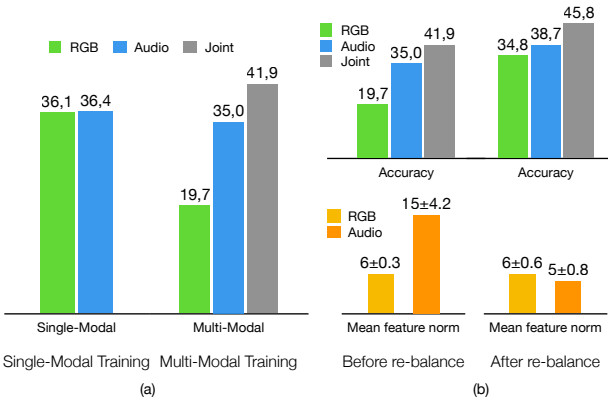

Figure 3. By jointly training, and testing on separate streams, the RGB performance drop (left). "unbalance" at feature-norm level which, when mitigated, leads to better performance (right).

Norm *unbalance*. We hypothesize that during training there is an "unbalance" between the two modalities that prevents the network from learning "equally" from the two. This hypothesis is also supported by the fact that the hyper-parameters discovered in [83] differ significantly depending on the modality. To empirically confirm this intuition, we performed a simple experiment, which is shown in Figure 3-a. Both modalities perform equally well at test time when RGB and audio streams are trained independently. However, when trained together and tested separately, the RGB accuracy decreases compared to the audio accuracy. This proves that the optimization of the RGB stream was negatively affected by multi-modal training. We also wondered whether this concept, i.e., the unbalance that occurs between modalities during the training phase, could be extended to a multi-source context. Is it possible that one source has a greater influence on the other, negatively affecting the final model? Based on the above considerations, we searched for a function that captures the amount of information contained in the final embedding of each modality, possibly justifying the existence of this unbalance.

The *mean feature norms*. Several works highlighted the existence of a strong correlation between the mean feature norms and the amount of "valuable" information for classification [92, 80, 62]. In particular, the cross-entropy loss has been shown to promote well-separated features with a high norm value in [80]. Moreover, the work of [89] is based on the Smaller-Norm-Less-Informative assumption, which implies that a modality representation with a smaller norm is less informative during inference. All of the above results suggest that the $L2$-norm of the features gives an indication of their information content, and thus can be used as

a metric to measure the unbalance between the two training modalities. By studying the behaviour of the feature norms, we found that, on the training set, the mean feature norms of audio samples ($\simeq 32$) was larger than that of RGB ones ($\simeq$10). This unbalance is also reflected on the test set (Figure 3-b, left), with the modality with the smaller norm being the one whose performance are negatively affected.

Motivated by these results, we propose a simple but effective loss whose goal is to re-balance the mean feature norms during training across multiple sources, so that the network can fully exploit joint training, especially in cross-domain scenarios. In fact, when re-balancing the norms, the performance of both modalities increase (Figure 3-b, right). Note that the concept of the smaller norm being less informative is used to argue that the network's preference for the audio modality is only due to its higher norm (over the RGB one), but this does not imply that RGB is less informative for the task; indeed, the range of norms after re-balancing is closer to the original RGB norm.

3.2. Domain Generalization

We assume to have different source domains $\{\mathcal{S}_1, ..., \mathcal{S}_k\}$, 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 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 main assumptions are that the distributions of all the domains are different, i.e., $\mathcal{D}_{s,k} \neq \mathcal{D}_t \wedge \mathcal{D}_{s,k} \neq \mathcal{D}_{s,j}$, with $k \neq j$, $k, j = 1, ..., k$, and that the label space is shared, $\mathcal{Y}_s = \mathcal{Y}_t$.

3.3. Framework

Let us consider an input $x = (x_i^v, x_i^a)$, where we denote with v and a the visual and audio modality respectively, and with i the i-th sample. As shown in Figure 5, each input modality (x_i^v, x_i^a) is fed to a separate feature extractor, F^v and F^a respectively. 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). They are then combined with a *late fusion* approach to obtain a final prediction (see Section 4 for more details). The whole architecture, which we call RNA-Net, is trained by minimizing the total loss, defined as

$$\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_{RNA}, \tag{1}$$

where the $\mathcal{L}_C$ is the standard *cross-entropy loss* and λ indicates the weight given to the proposed cross-modal loss, called Relative Norm Alignment loss ($\mathcal{L}_{RNA}$). Technical details of $\mathcal{L}_{RNA}$ are defined in the next section.

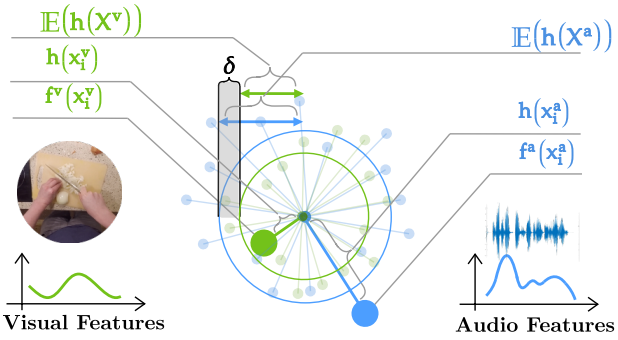

Figure 4. The norm $h(x_i^v)$ of the i-th visual sample (left) and $h(x_i^a)$ of the i-th audio sample (right) are represented, by means of segments of different lengths. The radius of the two circumferences represents the mean feature norm of the two modalities, and δ their discrepancy. By minimizing δ, audio and visual feature norms are induced to be the same.

3.4. Relative Norm Alignment Loss

Definition. The main idea behind our loss is the concept of *mean-feature-norm distance*. We denote with $h(x_i^m) = (\|\cdot\|_2 \circ f^m)(x_i^m)$ the L_2-norm of the features f^m of the m-th modality, and compute the *mean-feature-norm distance* (δ) between the two modality norms f^v and f^a as

$$\delta(h(x_i^v), h(x_i^a)) = |\mathbb{E}[h(X^v)] - \mathbb{E}[h(X^a)]| \tag{2}$$

where $\mathbb{E}[h(X^m)]$ corresponds to the mean features norm for each modality. Figure 4 illustrates the norm $h(x_i^v)$ of the i-th visual sample and $h(x_i^a)$ of the i-th audio sample, by means of segments of different lengths arranged in a radial pattern. The mean feature norm of the k-th modality is represented by the radius of the two circumferences, and δ is represented as their difference. The objective is to minimize the δ distance by means of a new loss function, which aims to align the mean feature norms of the two modalities. In other words, we restrict the features of both modalities to lie on a hypersphere of a fixed radius.

We propose a Relative Norm Alignment loss, defined as

$$\mathcal{L}_{RNA} = \left(\frac{\mathbb{E}[h(X^v)]}{\mathbb{E}[h(X^a)]} - 1 \right)^2, \tag{3}$$

where $\mathbb{E}[h(X^m)] = \frac{1}{N} \sum_{x_i^m \in \mathcal{X}^m} h(x_i^m)$ for the m-th modality and N denotes the number of samples of the set $\mathcal{X}^m = \{x_1^m, ..., x_N^m\}$. This dividend/divisor structure is introduced to encourage a relative adjustment between the norm of the two modalities, inducing an *optimal equilibrium* between the two embeddings. Furthermore, the square of the difference pushes the network to take larger steps when the ratio of the two modality norms is too far from one, resulting in faster convergence.

Conceptually, aligning the two modality norms corresponds to imposing a "hard" constraint, aligning them to a

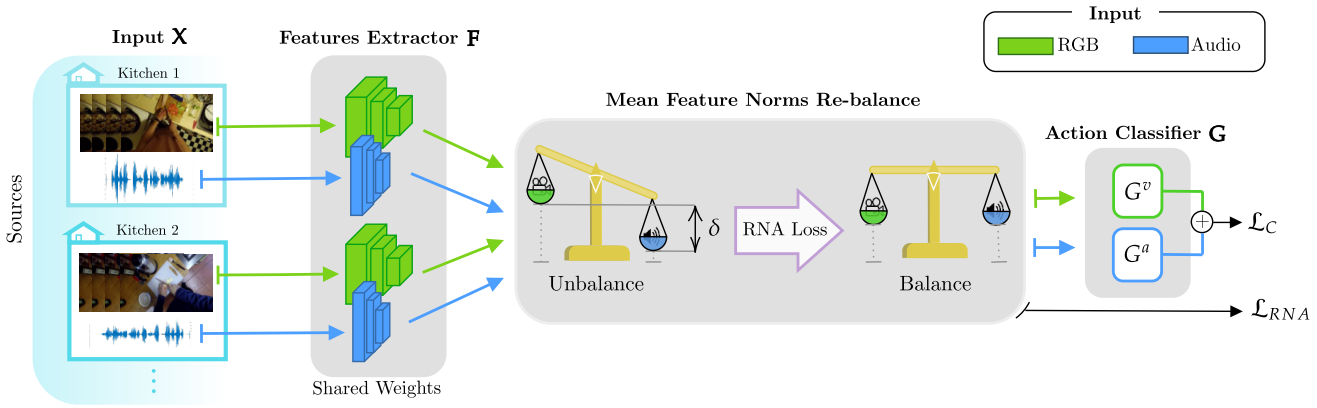

Figure 5. **RNA-Net.** Labeled source visual $x^v_{s,i}$ and source audio $x^a_{s,i}$ inputs are fed to the respective feature extractors F^v and F^a. Our loss $\mathcal{L}_{RNA}$ operates at the feature-level by balancing the relative feature norms of the two modalities. The network is trained with a standard cross-entropy loss $\mathcal{L}_c$ jointly with $\mathcal{L}_{RNA}$. At inference time, multi-modal target data is used for classification.

constant value k. We refer to this as *Hard Norm Alignment* (HNA), and we formulate the corresponding $\mathcal{L}_{HNA}$ loss as

$$\mathcal{L}_{HNA} = \sum_m \left(\mathbb{E}[h(X^m)] - k\right)^2, \tag{4}$$

where k is the same for all m modalities. Nevertheless, as shown in Section 4, our formulation of $\mathcal{L}_{RNA}$ helps the convergence of the two distributions' norms without requiring this additional k hyper-parameter. Designing the loss as a *subtraction* ($\mathcal{L}^{sub}_{RNA}$) between the two norms by directly minimizing δ^2 (Equation 2) (see Supplementary) is a more straightforward solution and a valid alternative. However, the rationale for this design choice is that a substantial discrepancy between the value of k and $\mathbb{E}[h(X^m)]$, as well as $\mathbb{E}[h(X^v)]$ and $\mathbb{E}[h(X^a)]$, would reflect in a high loss value, thus requiring an accurate tuning of the weights and consequently increasing the network sensitivity to loss weights [38]. Indeed, this dividend/divisor structure ensures that loss to be in the range $(0, 1]$, starting from the modality with higher norm as dividend.

***Learn* to re-balance.** The final objective of RNA loss is to *learn* how to leverage audio-visual norm correlation *at feature level* for a general and effective classification model. It is precisely because the network learns to solve this task that we obtain generalization benefits, rather than avoiding the norm unbalance through input level normalization or pre-processing. Note that introducing a normalization at *input level* could be potentially not suitable for pre-trained models. Moreover, it would not be feasible in DG, where the access to target data is not available during training, and thus not only there is no information on the target distribution, but each domain also requires a distinct normalization.

Additionally, the rationale behind *learning* to re-balance rather than using typical projection methods to normalize features [62] is two-fold. First, forcing the network to normalize the features using model weights themselves mitigates the "norm unbalance" problem also during inference, as the network has the chance to learn to work in the normalized feature space during training. Secondly, explicit normalization operators, e.g., batch normalization, impose the scaled normal distribution individually for each element in the feature. However, this does not ensure that overall mean feature norm of the two modalities to be the same.

3.5. Extension to Unsupervised Domain Adaptation

Thanks to the unsupervised nature of $\mathcal{L}_{RNA}$, our network can be easily extended to the UDA scenario. Under this setting, both labelled source data from a single source domain $\mathcal{S} = (\mathcal{S}^v, \mathcal{S}^a)$, and unlabelled target data $\mathcal{T} = (\mathcal{T}^v, \mathcal{T}^a)$ are available during training. We denote with $x_{s,i} = (x^v_{s,i}, x^a_{s,i})$ and $x_{t,i} = (x^v_{t,i}, x^a_{t,i})$ the i-th source and target samples respectively. Both $x^m_{s,i}$ and $x^m_{t,i}$ are fed to the feature extractor F^m of the m-th specific modality, shared between source and target, obtaining respectively the features $f_s = (f^v_s, f^a_s)$ and $f_t = (f^v_t, f^a_t)$. In order to consider the contribution of both source and target data during training, we redefine our $\mathcal{L}_{RNA}$ under the UDA setting as

$$\mathcal{L}_{RNA} = \mathcal{L}^s_{RNA} + \mathcal{L}^t_{RNA} \tag{5}$$

where $\mathcal{L}^s_{RNA}$ and $\mathcal{L}^t_{RNA}$ correspond to the RNA formulation in Equation 2, applied to source and target data respectively. Both the contributions are added in Equation 1.

4. Experiments

In this section, we first introduce the dataset used and the experimental setup (Section 4.1), then we present the experimental results (Section 4.2). We compare RNA-Net against existing multi-modal approaches, and both DG and UDA methods. We complete the section with some ablation studies and qualitative results.

4.1. Experimental Setting

Dataset. We use the EPIC-Kitchens-55 dataset [18] and we adopt the same experimental protocol of [56], where the three kitchens with the largest amount of labeled samples are handpicked from the 32 available. We refer to them here as D1, D2, and D3 respectively.

Input. Regarding the RGB input, a set of 16 frames, referred to as *segment*, is randomly sampled 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 [36] and convert the audio track into a 256×256 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.

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 [9] as done in [56]. The audio feature extractor uses the BN-Inception model [33] pretrained on ImageNet [22], which proved to be a reliable backbone for the processing of audio spectrograms [36]. 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 and the cross entropy loss is used to train the network. We train RNA-Net for $9k$ iterations using the SGD optimizer. The learning rate for RGB is set to $1e-3$ and reduced to $2e-4$ at step $3k$, while for audio, the learning rate is set to $1e-3$ and decremented by a factor of 10 at steps $\{1000, 2000, 3000\}$. The batch size is set to 128, and the weight λ of $\mathcal{L}_{RNA}$ is set to 1.

4.2. Results

DG Results. Table 1 illustrates the results of RNA-Net under the multi-source DG setting. We compare it to the *Deep All* approach, namely the backbone architecture when no other domain adaptive strategies are exploited and all the source domains are fed to the network. Indeed, this is the baseline in all image-based DG methods [7]. We adapted a well-established image-based domain generalization approach, namely IBN-Net [57], and the multi-modal self-supervised task proposed in MM-SADA [56] to evaluate the effectiveness of RNA-Net in the DG scenario. Indeed, training the network to solve a self-supervised task jointly with the classification has been demonstrated to be helpful in generalizing across domains [7]. Finally, we compare RNA-Net against Gradient Blending (GB) [83]. The results in Table 1 show that RNA-Net outperforms all the above mentioned methods by a significant margin.

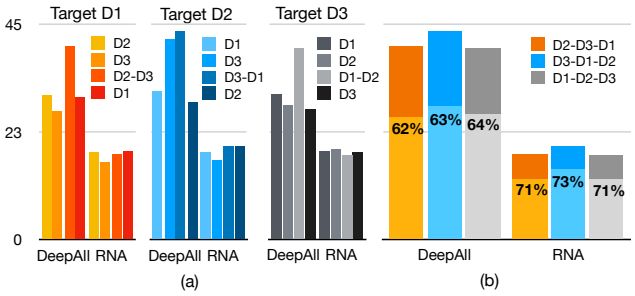

Figure 6. Mean feature norms of target when training on different source domains. The norm unbalance reflects also between different source domains, and RNA mitigates it (left). The percentage of the norm of the most relevant features over the total norm increases when minimizing RNA loss (right).

What are the benefits of RNA loss in generalizing? The RNA loss prevents the network from overspecializing across all domains from which it receives input. As it can be seen from Figure 6-a, the total norm (RGB+Audio) of target features varies greatly depending on the training source domains, increasing in the multi-source scenario (concept of "unbalance", Section 3.1). However, despite the lack of a direct constraint across the sources, by minimizing RNA loss together with the classification loss, the network learns a set of weights (shared across all sources) that re-balances the contribution of the various input sources. In such a way, the network can exploit the information from all the input sources equally, as it has been demonstrated that norm mismatch between domains account for their erratic discrimination [87]. We also noticed that by decreasing the total norm, the network promotes those features which are task-specific and meaningful to the final prediction, decreasing domain-specific ones which degrade performance on unseen data. This is shown in Figure 6-b, where we plot the total norm of the Top-300[1] features used for classification. By minimizing RNA loss, the percentage of this features have increased passing from up to 64% to up to 72% of the overall norm.

Best Single-Source. This experiment is a common practice in multi-source scenarios [81]. We choose the best source (the one with the highest performance) for each target, such as D2 for D3 (D2 $\rightarrow$ D3 $>$ D1 $\rightarrow$ D3) (Table 3). With this experiment, we aim to show that i) as a multi-modal problem, having many sources do not necessarily guarantee an improvement (DeepAll $<$ Best Single Source), therefore the need of using ad-hoc techniques to deal with multiple sources; ii) our loss allows the network to gain greater advantage from many sources (RNA-Net $>$ Best Single Source + RNA $>$ Best Single Source), confirming the domain generalization abilities of RNA-Net and the fact that it is not limited to tackle a multi-modal problem.

[1]The Top-300 is obtained selecting the features corresponding to the 300-weights of the main classifier that mostly affect the final prediction.

MULTI SOURCE DG				
	D2, D3 → D1	D3, D1 → D2	D1, D2 → D3	Mean
Deep All	43.19	39.35	51.47	44.67
IBN-Net [57]	44.46	49.21	48.97	47.55
MM-SADA (Only SS) [56]	39.79	52.73	51.87	48.13
Gradient Blending [83]	41.97	48.40	51.43	47.27
TBN [36]	42.35	47.45	49.20	46.33
Transformer [53]	42.78	47.38	51.79	47.32
Cross-Modal Transformer [14]	40.87	43.57	54.88	46.44
SE [32]	42.82	42.81	51.07	45.56
Non-Local [84]	45.72	43.08	49.49	46.10
RNA-Net (Ours)	45.65 ▲+2.46	51.64 ▲+12.32	55.88 ▲+4.41	**51.06** ▲+6.4

Table 1. Top-1 Accuracy (%) of RNA-Net in Multi Source DG scenario. In **green** we highlight improvement of RNA-Net w.r.t. the baseline Deep All.

UDA	
Method	Mean
Source-Only	41.87
MM-SADA (Only SS) [56]	46.44
GRL [28]	43.67
MMD [51]	44.46
AdaBN [48]	41.92
RNA-Net (Ours)	**47.71**
MM-SADA (SS+GRL) [56]	47.75
RNA-Net+GRL (Ours)	**48.30**

Table 2. Top-1 Accuracy (%) of RNA-Net in UDA context.

	Target: D1		Target: D2		Target: D3						
	D2 → D1	D3 → D1	D1 → D2	D3 → D2	D1 → D3	D2 → D3	Mean	D2, D3 → D1	D3, D1 → D2	D1, D2 → D3	Mean
Baseline (RGB Only)	34.76	33.03	34.15	41.08	35.03	38.79	36.14	39.72	33.59	37.91	37.07
Baseline (Audio Only)	29.17	34.00	31.90	42.93	36.49	44.00	36.42	41.57	39.21	47.19	42.66
Baseline	35.27	<u>40.26</u>	39.03	<u>49.98</u>	39.17	<u>47.52</u>	41.87	43.19	39.35	51.47	44.67
RNA-Net	41.76	<u>42.20</u>	45.01	<u>51.98</u>	44.62	<u>48.90</u>	**45.75**	45.65	51.64	55.88	**51.06**
Best Single-Source	✗	Best D1	✗	Best D2	✗	Best D3		40.26	49.98	47.52	45.92
Best Single-Source + RNA	✗	Best D1	✗	Best D2	✗	Best D3		42.20	51.98	48.90	47.69

Table 3. Top-1 Accuracy (%) of RNA-Net w.r.t. uni-modal, multi-modal baseline and the *Best Single-Source* both w/o and w RNA loss. **Bold**: highest mean result, <u>underline</u> the best Single-Source case.

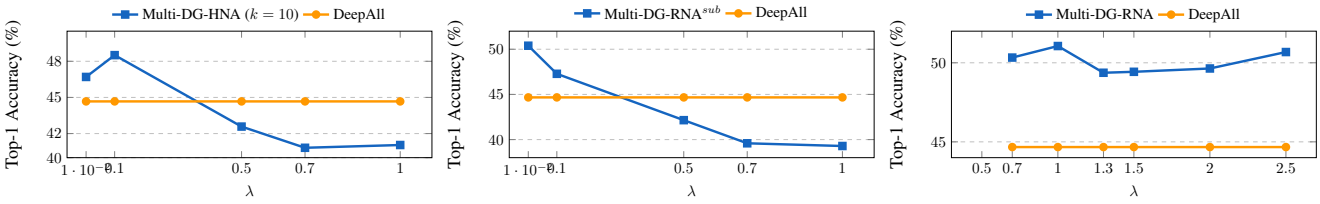

Figure 7. Different performance (average Top-1 Accuracy (%)) based on the value of λ used to weight HNA and RNA^{sub} and RNA losses.

ABLATION STUDY				
	D2, D3 → D1	D1, D3 → D2	D1, D2 → D3	Mean
DeepAll	39.35	43.19	51.47	44.67
HNA	49.41	45.88	50.24	48.51
RNA^{sub}	49.97	47.88	53.33	50.39
RNA	51.64	45.65	55.88	**51.06**
Supervised	50.11	63.62	65.56	59.76
Supervised +RNA	54.68	67.48	67.24	**63.13**

Table 4. Accuracy (%) of HNA, RNA^{sub} and RNA losses proposed in the main paper. In **bold** we show the highest mean result.

Multi-Modal Approaches. In Table 1 we compare RNA-Net against recent audio-visual methods, which we adapted to our setting. This is to verify if increasing cooperation between the two modalities, through other strategies, still improves the network's generalization abilities. Those are TBN [36], based on temporal aggregation, and two multi-modal Transformer-based approaches [14, 53]. Finally, since gating fusion approaches have been demonstrated to be valid multi-modal fusion strategies [39], we adapted Squeeze And Excitation [32] and Non-Local [84]. We leave details about the implementation in the Supplementary. These experiments confirm that by enhancing the audio-visual modality's cooperation the network's generalization improves. RNA-Net, on the other hand, surpasses all of those approaches by a large margin, demonstrating yet again how useful it is in cross-domain scenarios.

DA Results. Results in UDA, when target data (unlabeled) is available at training time, are summarized in Table 2. We validate RNA-Net against four existing UDA approaches, namely AdaBN [48], MMD [51], GRL [28] and MM-SADA[2] [56]. The baseline is the *Source-Only* (training on source and testing directly on target data). MM-SADA [56] combines a self-supervised approach (SS) with an adversarial one (GRL). We compare RNA-Net with both the complete method and its DA single components (SS, GRL). Interestingly, it provides comparable results to MM-SADA despite not being expressly designed as a UDA-based technique. It should also be noted that MM-SADA

[2]To put MM-SADA [56] on equal footing to RNA-Net, we run it with audio-visual input

Method	S-DG	M-DG	Norm	Angle
DeepAll	41.87	44.67		
CosSim	41.76	45.60		✓
MSE	45.52	46.11	✓	✓
Orth.Loss	42.67	47.64		✓
RNA loss	**45.75**	**51.06**	✓	

Table 5. Comparison in terms of accuracy (%) between RNA loss and other existing losses.

EPICKITCHEN-100			
	Target	Top-1	Top-5
Source Only	✗	44.39	69.69
TA^{3}N [12]	✓	46.91	72.70
RNA-Net	✗	47.96	79.54
TA^{3}N [12]+RNA-Net	✓	**50.40**	**80.47**

Table 6. Verb accuracy (%) on the EK-100 UDA setting.

must be trained in stages, while RNA-Net is end-to-end trainable. Finally, we prove the complementarity of our approach with the adversarial one (RNA-Net+GRL), achieving a slight improvement over MM-SADA.

Ablation Study. In Table 3 we show the performance of the two modalities when trained separately (*RGB Only*, *Audio Only*) and tested directly on unseen data, showing that the fusion of the two streams provides better results. In Table 4 we also perform an ablation study to validate the choices on the formulation of RNA loss. In particular, we compare it against the *Hard Norm Alignment* loss (HNA), and against RNA^{sub}, confirming that the dividend/divisor structure is the one achieving better performance. Finally, we show that our loss not only benefits across domains, but also improves performance in the supervised setting.

Ablation on λ variations. In Figure 7, we compare the performances of HNA, RNA^{sub}, and RNA respectively as a function of λ. Results show that the performance of both HNA and RNA^{sub} are highly sensitive to λ. Specifically, higher values of λ cause significant performance degradations since (potential) large difference values between $\mathbb{E}[h(X^a)]$ and $\mathbb{E}[h(X^v)]$ (for RNA^{sub}) or k and $\mathbb{E}[h(X^m)]$ (for HNA) result in high loss values that could cause the network to diverge. These convergence problems are softened by the "ratio" structure of RNA, which outperforms the baseline results on all choices of λ.

Comparison with other losses. We compare the RNA loss against a standard cosine similarity-based loss and an Euclidean-based loss, i.e., the Mean Square Error (MSE) (Table 5). The first enforces alignment by minimizing the angular distance between the two feature representations, and the second tends to align both their angular and norms by minimizing the $L2$-loss of the two. The results suggest that re-balancing the norms has a greater impact than not using angular limitations. In fact, RNA (norm re-balance, no angular constraint) outperforms MSE (both norm re-balance and angular constraint), notably in multi-DG. We believe that a loss should allow feature distributions to retain modality-specific features when one modality is weak or contains only domain-information, and to align them when both are connected with action. To this purpose, we compare RNA loss to an orthogonality loss, which keeps modality-specific properties rather than aligning them (details on Supplementary). As shown in Table 5, the Orth. loss outperforms the CosSim (alignment) loss, especially in multi-DG, proving that utilizing modality-specific features deals with domain shift better. The RNA loss outperforms it by a significant margin. Our intuition is that by not constraining the angle, we do not strictly enforce an alignment (when non-necessary) or an orthogonality, letting the network to find itself the better angle for the main task.

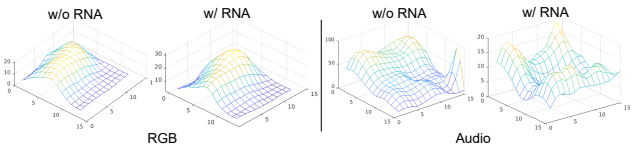

Figure 8. Norm ranges of RGB and audio. When minimizing the RNA loss, relevant features are kept high, while the less relevant are decreased (see yellow peaks). Best viewed in colors.

EPIC-Kitchens-100 UDA. The results achieved on the recently proposed EPIC-Kitchens-100 UDA setting [19] are shown in Table 6. The source is composed of videos from EK-55, while the target is made up of videos from EK-100. RNA-Net outperforms the baseline Source Only by up to 3% on Top-1 and 10% on Top-5, remarking the importance of using ad-hoc techniques to deal with multiple sources (see Supplementary). Moreover, it outperforms the very recent UDA technique TA^{3}N [12] ***without access to target data***. Interestingly, when combined to TA^{3}N, it further improves performance, proving the complementarity of RNA-Net to other existing UDA approaches.

Qualitative Analysis. In Figure 8 we empirically show the effect of RNA loss on feature norms, by analyzing the behaviour of the spatial features' norms, i.e., features before the last average pooling. When increasing the mean feature norms (in the case of RGB), the most significant features are increased, while when decreasing them (in the case of audio) the unrelevant (domain-specific) ones are reduced. This is evident especially in the case of audio (right).

5. Conclusion

In this paper we showed for the first time that generalization to unseen domains in first person action recognition can be achieved effectively by leveraging over the complementary nature of audio and visual modalities, bringing to light the "norm unbalance" problem. We presented an innovative vision on multi-modal research, proposing the modality feature norms as a measure unit. To this end, we designed a new cross-modal loss that operates directly on the relative feature norm of the two modalities. We see this norm-based approach as a promising new take on multi-modal learning, potentially of interest for many other research fields.

Acknowledgements. The work was partially supported by the ERC project N. 637076 RoboExNovo and the research herein was carried out using the IIT HPC infrastructure. This work was supported by the CINI Consortium through the VIDESEC project.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020.

[2] Nakul Agarwal, Yi-Ting Chen, Behzad Dariush, and Ming-Hsuan Yang. Unsupervised domain adaptation for spatio-temporal action localization. *arXiv preprint arXiv:2010.09211*, 2020.

[3] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019.

[4] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019.

[5] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[6] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[7] Silvia Bucci, Antonio D'Innocente, Yujun Liao, Fabio Maria Carlucci, Barbara Caputo, and Tatiana Tommasi. Self-supervised learning across domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[8] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[9] 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.

[10] Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, and Mariella Dimiccoli. Seeing and hearing egocentric actions: How much can we learn? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[11] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019.

[12] 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.

[13] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020.

[14] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3884–3892, 2020.

[15] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020.

[16] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016.

[17] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9944–9953, 2019.

[24] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32:6450–6461, 2019.

[25] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011.

[26] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[27] Antonino Furnari and Giovanni Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[28] 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.

[29] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020.

[30] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019.

[31] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[33] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. pages 448–456, 2015.

[34] A. Jamal, Vinay P. Namboodiri, Dipti Deodhare, and K. Venkatesh. Deep domain adaptation in action space. In *BMVC*, 2018.

[35] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, and Remco Veltkamp. Multitask learning to improve egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

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

[37] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859. IEEE, 2021.

[38] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[39] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[40] Bruno Korbar. Co-training of audio and video representations from self-supervised temporal synchronization. 2018.

[41] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7774–7785. Curran Associates Inc., 2018.

[42] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018.

[43] Nallapaneni Manoj Kumar, Neeraj Kumar Singh, and VK Peddiny. Wearable smart glass: Features, applications, current progress and challenges. In *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 577–582. IEEE, 2018.

[44] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion feature network: Fixed motion filter for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–403, 2018.

[45] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[46] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, 2018.

[47] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[48] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

[49] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[50] 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.

[51] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[52] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016.

[53] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. *arXiv preprint arXiv:2011.01819*, 2020.

[54] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12934–12945, June 2021.

[55] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

[56] 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.

[57] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. *arXiv preprint arXiv:1805.07925*, 2018.

[58] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[59] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11815–11822, 2020.

[60] 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.

[61] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*, 2018.

[62] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[63] Ivan Rodin, Antonino Furnari, Dimitrios Mavroedis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, page 103252, 2021.

[64] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[65] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[66] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 568–576, Cambridge, MA, USA, 2014. MIT Press.

[67] Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2620–2628, 2016.

[68] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9787–9795, June 2021.

[69] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019.

[70] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1111, 2020.

[71] Swathikiran Sudhakaran and Oswald Lanz. Convolutional long short-term memory networks for recognizing first person interactions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[72] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. *arXiv preprint arXiv:1807.11794*, 2018.

[73] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

[74] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *AAAI*, pages 5940–5947, 2020.

[75] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: weakly-supervised audio-visual video parsing. *arXiv preprint arXiv:2007.10558*, 2020.

[76] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[77] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[78] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[79] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems*, pages 5334–5344, 2018.

[80] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.

[81] Hang Wang, Minghao Xu, Bingbing Ni, and Wenjun Zhang. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In *European Conference on Computer Vision*, pages 727–744. Springer, 2020.

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

[83] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020.

[84] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[85] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[86] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[87] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1435, 2019.

[88] Zhiyu Yao, Yunbo Wang, Xingqiang Du, Mingsheng Long, and Jianmin Wang. Adversarial pyramid network for video domain generalization. *arXiv preprint arXiv:1912.03716*, 2019.

[89] Jianbo Ye, Xin Lu, Zhe Lin, and James Z Wang. Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers. *ICLR*, 2018.

[90] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[91] Jiaojiao Zhao and Cees GM Snoek. Dance with flow: Two-in-one stream action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9935–9944, 2019.

[92] Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5089–5097, 2018.

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