Relative Norm Alignment for Tackling Domain Shift in Deep Multi-modal Classification

Relative Norm Alignment for Tackling Domain Shift in Deep Multi-modal Classification

Mirco Planamente[1,2,3*], Chiara Plizzari[1], Simone Alberto Peirone[1], Barbara Caputo[1,3], Andrea Bottino[1]

[1]DAUIN, Politecnico di Torino, Torino, Italy.
[2]Italian Institute of Technology, Genova, Italy.
[3]Consortium Cini, Italy.

*Corresponding author(s). E-mail(s): mirco.planamente@polito.it;
Contributing authors: chiara.plizzari@polito.it; simone.peirone@polito.it; barbara.caputo@polito.it; andrea.bottino@polito.it;

Abstract

Multi-modal learning has gained significant attention due to its ability to enhance machine learning algorithms. However, it brings challenges related to modality heterogeneity and domain shift. In this work, we address these challenges by proposing a new approach called Relative Norm Alignment (RNA) loss. RNA loss exploits the observation that variations in marginal distributions between modalities manifest as discrepancies in their mean feature norms, and rebalances feature norms across domains, modalities, and classes. This rebalancing improves the accuracy of models on test data from unseen ("target") distributions. In the context of Unsupervised Domain Adaptation (UDA), we use unlabeled target data to enhance feature transferability. We achieve this by combining RNA loss with an adversarial domain loss and an Information Maximization term that regularizes predictions on target data. We present a comprehensive analysis and ablation of our method for both Domain Generalization and UDA settings, testing our approach on different modalities for tasks such as first and third person action recognition, object recognition, and fatigue detection. Experimental results show that our approach achieves competitive or state-of-the-art performance on the proposed benchmarks, showing the versatility and effectiveness of our method in a wide range of applications.

Keywords: Multi-modal learning, Norm alignment, Domain Generalization, Unsupervised Domain Adaptation

1 Introduction

Humans have the ability to perceive the world around them through signals that come from multiple sensory systems. Our perceptual experiences can be visual, auditory, tactile, olfactory, and gustatory. Psychologists and neurologists agree that our perception does not depend on a single modality at a single time, but is fundamentally multi-modal in nature [1, 2]. Moreover, the interpretation of data from one sensory channel is influenced by data from other modalities [3, 4].

The same ability to effectively process and integrate information from multiple sensory channels has been shown to significantly improve the performance of current machine learning algorithms. For example, recent video understanding

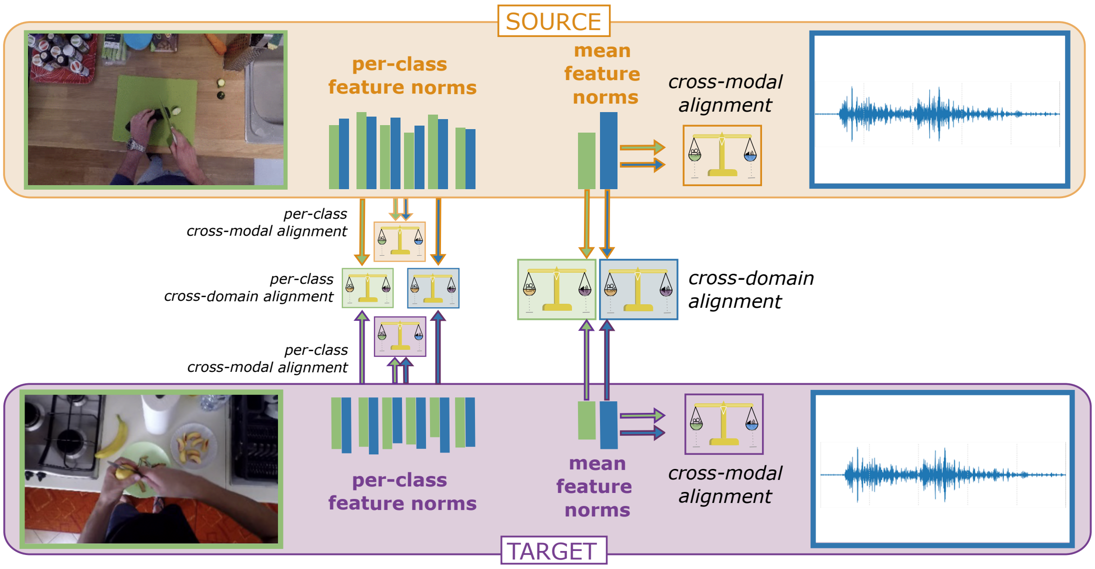


Fig. 1 Overview of Relative Norm Alignment (RNA) loss for RGB and audio modalities. Given visual and audio input from both source and target domains, we perform an alignment at feature level by re-balancing (i) the mean feature norms of visual and audio modalities (*cross-modal alignment*, $\mathcal{L}^{g}_{RNA}$), (ii) per-class mean feature norms of visual and audio modalities (*per-class alignment*, $\mathcal{L}^{c}_{RNA}$) and (iii) mean feature norms of source and target features independently for each modality (*cross-domain alignment*, $\mathcal{L}^{mod}_{RNA}$).

models use complementary audio-visual [5] and appearance-motion information [6–8] to improve accuracy and generalization performance. Object recognition algorithms use depth information to extract more accurate information and classify objects more effectively [9, 10], and so on.

Despite its potential benefits, Multi-Modal Learning (MML) also comes with some challenges, such as learning how to summarize data while retaining their complementary information [11] or understanding how to effectively combine information from multiple modalities when making a prediction [12]. The same issue is addressed in [13] when data from multiple views are used. Heterogeneity between modalities is another critical issue, as the difference between their marginal distributions may prevent the model from learning equally from all of them [14]. Another well-known problem in the literature is the so-called "domain shift", i.e., that a model trained on a labeled source dataset does not generalize well to an unseen target dataset. Different modalities may be affected differently by the domain shift [15]. For example, when using audio-visual data for egocentric action recognition, the action "cut" in a cooking scenario may reveal differences between domains [16], as cutting boards in different kitchens may differ in their visual and auditory impressions (e.g., wooden cutting board vs. plastic cutting board), different types of food may be cut, and so on. This highlights the need for robust models that can handle variation across modalities and domains.

To address both the cross-modal and cross-domain challenges in MML, we recently proposed in [16] a simple but effective audio-visual approach in the context of egocentric action recognition. We observed that differences in the marginal distributions of the audio and visual modalities could lead to variations in feature informativeness that do not only negatively affect the training process and lead to suboptimal performance, but also typically translate into discrepancies between the mean norms of their features. This imbalance in norms leads the network to "favor" the modality with the larger features, which prevents the model from fully exploiting the synergies and complementarities between modalities and reduces its generalization capabilities [17].

To tackle this issue, in [16] we proposed to reduce such imbalance with a simple loss called *Relative Norm Alignment* (RNA) loss. In the Domain Generalization (DG) setting, i.e., when the model does not have access to the target data at training time, this loss attempts to align the average norms of the different modalities to a common value. This objective also leads to successful transfer between source and target [17–20]. In the Unsupervised Domain Adaptation (UDA) setting, i.e., when target data are available during training, RNA is defined as the sum of two domain-specific terms that aim to achieve a cross-modality norm balance on both source and target domains. However, in this setting, the RNA loss operated separately on the two domains (*cross-modal alignment*, Fig. 1), resulting in discrepancies between the mean feature norms of the two. This discrepancy can be explained by the presence of domain-specific features from the source domain that may have low activations in the target domain.

To improve the effectiveness of RNA, in this work we extend it to independently align feature norms for each modality across domains (*cross-modal alignment*, Fig. 1) so that the network can prioritize more transferable features [20]. In addition, we address the problem of imbalanced feature norms between classes by introducing an intra- and inter-domain alignment component per class (*per-class alignment*, Fig. 1), resulting in improved overall accuracy.

Furthermore, we combine RNA with two additional components in UDA settings. First, we incorporate an adversarial loss to improve domain-invariant feature learning. Second, we observe that the original RNA loss only affects the modality embedding models and neglects the classification layers. To mitigate the prediction uncertainty in the target domain, we extend the training loss of the model with an Information Maximization term that uses pseudo-labels on target data.

The solution we propose differs from previous approaches in that it is simple, it does not require changes to the training process and, differently from recent constrastive-learning based approaches [21–23], it does not require effective mining of hard negative samples. This makes our solution a desirable choice for a broader range of modalities and tasks. In particular, we extend the audio-visual loss proposed in [16] to a variety of visual and nonvisual modalities (optical flow, event data, depth, EGG, facial keypoints) and to a variety of tasks, including first- and third-person action recognition, object recognition, and fatigue detection. Despite its simplicity, experiments show that our approach performs equally well, if not better, than existing methods, with a leaner and more efficient implementation.

In summary, the main contributions of this work are as follows:

- it updates the definition of RNA to improve the transferability of features between domains in DG and UDA settings;
- it introduces the use of pseudo-labeling to regularize predictions in the context of transfer learning between source and target;
- it addresses the challenges of multi-modal domain shift by extending our analysis to multiple modalities and multiple tasks;
- it presents a comprehensive analysis and ablation of our approach in both DG and UDA settings, showing state-of-the-art or competitive performances on all benchmarks.

2 Related Work

MML has gained popularity due to its potential for better performance, robustness, and deeper understanding. Previous surveys [12, 24] have discussed the challenges and opportunities of MML. Here we focus on the problem of generalizing MML across domains, and in particular explore computer vision applications that are most relevant to the experiments in our work.

Domain Adaptation. In DG, the goal is to build a model that uses knowledge from one or multiple source domains, without having access to data from the target domain during training, to improve the generalization performance of the model to any unseen domain. In such a setting, the lack of knowledge about the target distributions prevents the possibility to estimate the domain discrepancy between source and target domains. Computer vision based DG approaches have mainly focused on image data and can be broadly classified into several categories. Feature-based methods aim to learn domain-invariant representations by aligning domain distributions with metrics such as MMD [25, 26] or CORAL [27], or with domain adversarial networks [28]. Data-based methods increase the amount of training data to prevent overfitting or use style transfer to reduce the domain sensitivity [29–34]. Meta-Learning methods simulate the shift in distributions between domains [35–37]. Self-Supervision [38] uses auxiliary tasks to learn generalizable representations.

In the context of video data, VideoDG [39] observes that it is important to find a balance between the ability to generalize and the ability to discriminate. To achieve this, the relationships between frames in the source domain are extended to ensure that they can generalize to potential target domains while maintaining their discriminative capabilities.

As for UDA methods (which can benefit from unlabeled target data available during training), they can be broadly divided into two categories: discrepancy-based and adversarial-based methods. Discrepancy-based methods minimize a distance metric between the source and target distributions [20, 40, 41]. Adversarial-based methods, on the other hand, use adversarial training to align source and target distributions [42, 43]. Another research direction focuses on incorporating self-supervised learning as an auxiliary task to improve feature learning, as in [38].

While the aforementioned approaches have mainly been applied to standard image classification tasks, there has also been a significant amount of research on UDA for video-related tasks, such as action detection [44], segmentation [45], and classification [6, 22, 46–49].

In video classification, several methods have been proposed to align the temporal dynamics of the feature space. TA^{3}N [46] uses a multi-level adversarial framework with temporal relation and attention mechanisms to achieve this goal. TCoN [49] aligns feature distributions between source and target domains with a cross-domain co-attention mechanism that focuses on aligning temporal relationship features to increase robustness across domains. In [47], the network is trained to solve an auxiliary self-supervised task on source and target data. SAVA [50] addresses the domain adaptation problem by proposing to use clip order prediction as an auxiliary task to be solved in both source and target domains. In addition, Contrastive Learning (CL) methods have also been proposed for UDA in video analysis. For example, CoMix [21] introduced a new framework for contrastive learning that aims to learn discriminative invariant feature representations.

Multi-Modal Adaptation and Generalization. Several methods have been proposed to exploit the availability of multiple modalities for domain adaptation. They can be divided into three main categories: adversarial approaches, co-training, and contrastive learning based methods.

Adversarial-based approaches, such as MDANN [51] and AUDA [52], focus on learning discriminative and domain adaptive features under an adversarial objective, showing their effectiveness in cross-domain emotion recognition using audio-visual data and cross-media retrieval using images and text from different domains. MM-SADA [6] is another approach that extends adversarial alignment to a self-supervised task based on modality correspondence.

Co-training methods such as DLMM [15] and XM-UDA [53] exploit the diverse properties of the different modalities by treating the classifiers of the various modalities as a set of teacher/student models trained with a curriculum learning approach. These methods have been applied to tasks such as event recognition using audio-visual data, fatigue detection using EEG signals and facial keypoints, and action recognition using RGB images and optical flow.

Contrastive learning based methods such as STCDA [22] and the approach described in [23] exploit the complementarity of different modalities to regularize both cross-modal and cross-domain feature representations. They treat each modality as a view and perform contrastive learning across modalities and domains to align representations between source and target domains in each modality. CIA [54] uses cross-modal interaction and generative modelling to align cross-domain representations.

RNA-Net [16] addresses multi-modal video DG by using both audio and RGB features, but recognizes that the simple fusion of multi-modal information may not improve generalizability. To overcome this problem, a cross-modal audio-visual Relative Norm Alignment (RNA) loss is proposed to align the relative feature norms of audio and visual modalities from source domains, resulting in domain-invariant audio-visual features. In this work, we further extend this approach to improve feature transferability across domains in both UDA and DG settings, and address the issues of multi-modal domain shift across different tasks and datasets.

Norm Alignment. Several works highlighted the existence of a strong correlation between the mean feature norms and the amount of "valuable" information for classification [19, 55, 56] and the negative impact of different feature norms on multiview clustering approaches [57]. In particular, the cross-entropy loss has been shown to promote well-separated features with a high norm value [55]. Starting from this observation, the authors of [20] show that the main reason behind performance degradation on unseen data is the reduction in feature norms compared to the source domain. This stems from the fact that the supervision on the source domain causes the classifier to rely on domain-specific features that may not be present in the target domain, thus reducing the activations in the representation of the target features and consequently the norms of the target features. To address this problem, [20] introduced a loss that forces the norms between the two domains to adapt to increasingly larger scalars, resulting in improved transfer between domains.

Similarly, [17] proposed a regularization objective that promotes uniform feature norms between source and target representations while also inducing progressively higher norm values. Furthermore, they introduced an inter-class norm alignment objective, based on the observation that classes with higher confidence are associated with larger feature norms, to soften distribution biases towards the most frequent classes, whose higher classification confidence is typically associated with larger feature norms.

Subsequent works have demonstrated the effectiveness of incorporating this regularization term into various approaches to learn domain-invariant features, such as the adversarial distribution adaptation network proposed in [18] and the hierarchical transfer network described in [58].

In this work, we apply the concept of norm alignment to domain adaptation by extending it to a multi-modal setting, where the alignment is performed not only between domains, but also across modalities and classes. This allows us to better handle the complexity of multi-modal data and improve the transferability of features across different domains and modalities.

3 Proposed method

In the following, we detail the proposed Relative Norm Alignment (RNA) loss, which aims to mitigate the domain shift in MML by aligning the mean feature norms from different modalities (*cross-modal alignment*) and from different domains (*cross-domain alignment*), both globally and at class level.

3.1 Intuition and motivation

Joint training of multi-modal models may result in sub-optimal synergies between the different modalities. This observation has been theoretically demonstrated in [59], showing that naive joint training prevents efficient learning from all modalities. From an optimization perspective, the modality with better performance contributes to lower joint discriminative loss and dominates the training progress, while smaller gradient updates are propagated through the other modalities, leading to an under-optimized situation in which the dominant modality learn faster than the others [60]. In turn, the cross entropy loss encourages the network to learn more separable features [61], thus increasing their feature norms unevenly. This problem becomes particularly relevant in cross-domain scenarios, where the accuracy drop is further exacerbated by domain shift.

For this reason, in this work we introduce a new loss function based on the mean features norm of the different modalities. This loss promotes balanced learning and synergistic integration of modalities. By addressing the issues of modality imbalances and domain shift, RNA improves the model's ability to effectively exploit multi-modal information and improve overall performance.

3.2 Setting

Suppose we observe data $\mathcal{X}_\mathcal{S} = \{(x_{s,i}, y_{s,i})\}_{i=1}^{n_s}$ from a source distribution $\mathcal{S}$, where n_s is the total number of samples, each associated with a label $y_{s,i}$ from the label space $\mathcal{Y}_s$. Each sample $x_{s,i}$ contains multiple modalities, i.e., $x_{s,i} = \{x_{s,i}^1, \ldots, x_{s,i}^M\}$, where $x_{s,i}^m$ denotes the m^{th} modality of the i^{th} sample and M is the number of modalities. The target domain $\mathcal{T}$ comprises n_t annotated target samples $\mathcal{X}_\mathcal{T} = \{x_{t,i}\}_{i=1}^{n_t}$, each characterized by the same M modalities of the source samples (i.e., $x_{t,i} = \{x_{t,i}^1, \ldots, x_{t,i}^M\}$).

We assume that the distributions of all involved domains are different, i.e., $\mathcal{D}_{d_1}^j \neq \mathcal{D}_{d_2}^k$, where d_1 and d_2 are the domains (source or target) and j and k represent different modalities on

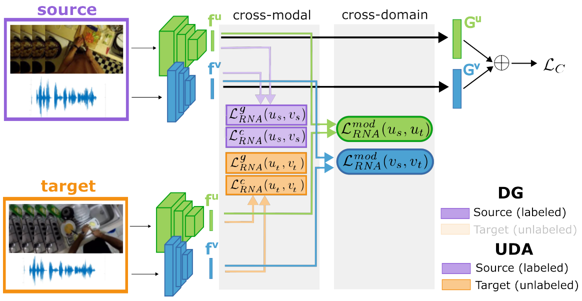

source
target
f^u
f^v
f^u
f^v
cross-modal
cross-domain
$\mathcal{L}^g_{RNA}(u_s, v_s)$
$\mathcal{L}^c_{RNA}(u_s, v_s)$
$\mathcal{L}^g_{RNA}(u_t, v_t)$
$\mathcal{L}^c_{RNA}(u_t, v_t)$
$\mathcal{L}^{mod}_{RNA}(u_s, u_t)$
$\mathcal{L}^{mod}_{RNA}(v_s, v_t)$
G^u
G^v
$\mathcal{L}_C$
DG
Source (labeled)
Target (unlabeled)
UDA
Source (labeled)
Target (unlabeled)

Fig. 2 Labeled source and unlabeled target samples from the modalities u (e.g., visual) and v (e.g., audio) are fed to the respective feature extractors. $\mathcal{L}_{RNA}$ aims to balance the relative feature norms of the two modalities, through a combination of the (domain-specific) cross-modal components ($\mathcal{L}^g_{RNA}$ and $\mathcal{L}^c_{RNA}$) and the cross-domain ones ($\mathcal{L}^{mod}_{RNA}$) in each u and v modality. In DG, only the components computed on the source are used.

the same domain and the same or different modalities on different domains. We also assume that the label space is shared between sources and targets, i.e., $\mathcal{Y}_s = \mathcal{Y}_t$.

3.3 RNA for Domain Adaptation

In the following, without loss of generality, we consider a single-source single-target problem in which two modalities are available. In Sec. 3.5 we show how the approach can be extended to any number of modalities.

We denote each input sample i as $x_i = (x_i^u, x_i^v)$, where u and v represent the two modalities (e.g., visual and audio modality). As shown in Fig. 2, each input modality m is fed to a separate features extractor F^m. The features $f_i^m = F^m(x_i^m)$ are then processed by a classifier G^m, which outputs the score predictions for the m^{th} modality of the i^{th} sample. Finally, the prediction scores from all modalities are combined using a *late fusion* approach to obtain the final predictions. In UDA settings, the F^m feature extractors are shared between source and target.

As previously mentioned, in this work we extend the approach introduced in [16], which proposes to train the entire architecture by minimizing the following loss:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}^g_{RNA}$$

where $\mathcal{L}_C$ is the standard *cross-entropy loss* on source data. The latter aims at *globally* minimizing the difference between the feature norms of the two modalities and is defined as:

$$\mathcal{L}^g_{RNA}(u, v) = \lambda_g \left(\frac{\mathbb{E}[h(X^u)]}{\mathbb{E}[h(X^v)]} - \frac{\mathbb{E}[h(X^v)]}{\mathbb{E}[h(X^u)]} \right)^2 \quad (1)$$

where $h(x_i^m) = (\|\cdot\|_2 \circ F^m)(x_i^m)$ is the L_2-norm of m^{th} modality features of the i^{th} sample, $\mathbb{E}[h(X^m)] = 1/B \sum_{x_i^m \in \mathcal{X}^m} h(x_i^m)$ is the average norm for the m^{th} modality of the B samples composing the batch, and λ_g weights $\mathcal{L}^g_{RNA}$. To ensure that all features have the same dimension, we project them to a common shape using a fully connected layer when this condition is not met.

In DG, the RNA objective is defined as $\mathcal{L}_{RNA} = \mathcal{L}^g_{RNA}(\mathcal{S})$ while in UDA $\mathcal{L}_{RNA} = \mathcal{L}^g_{RNA}(\mathcal{S}) + \mathcal{L}^g_{RNA}(\mathcal{T})$, where $\mathcal{L}^g_{RNA}(\mathcal{S})$ and $\mathcal{L}^g_{RNA}(\mathcal{T})$ are the loss in Eq. 1 applied to the source and target domains, respectively.

The dividend/divisor structure of $\mathcal{L}^g_{RNA}$ promotes a relative adjustment between the global norm of the two modalities aimed at achieving an *optimal equilibrium* between the two. The square of the difference forces the network to take larger steps when the ratio of the two modality norms is too different, leading to faster convergence. We note that Eq. 1 redefines the loss presented in [16] to ensure a symmetric form (i.e., $\mathcal{L}^g_{RNA}(u, v) = \mathcal{L}^g_{RNA}(v, u)$).

3.4 RNA extensions

While the results in [16] show the effectiveness of $\mathcal{L}^g_{RNA}$ in reducing domain shifts, the formulation in Eq. 1 has two major limitations. First, the *global* cross-modal alignment performed by $\mathcal{L}^g_{RNA}$ may also lead to unbalanced norms between modalities at the class level, which in turn tends to favor one modality over the others when making decisions about particular classes. Second, in UDA, the alignment is performed separately for each domain. As a result, the average feature norms may still show large differences between source and target domains. These differences can be attributed to the presence of domain-specific features that originate from training in the source domain and may have low activations in the target domain [17, 20], affecting overall accuracy.

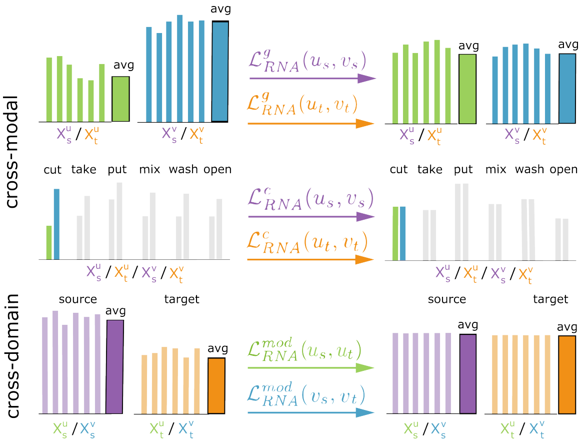


Fig. 3 *Individual effects* of $\mathcal{L}_{RNA}$ components on feature norms. Each diagram shows norms per class for a single modality and domain (u or v for source or target). **1st row:** $\mathcal{L}^g_{RNA}$ minimizes overall average norms (larger bars on the right) of u and v modalities. **2nd row:** $\mathcal{L}^c_{RNA}$ achieves balanced norms at class level. **3rd row:** $\mathcal{L}^{mod}_{RNA}$ balances class and average norms of the same modality across domains. The diagrams show the norms before (left) and after (right) applying the corresponding $\mathcal{L}_{RNA}$ component.

To address both problems, we propose the following extensions to the RNA formulation. First, we introduce an intra-domain class constraint $\mathcal{L}^c_{RNA}$ to address the cross-modal norm imbalance at class level, defined as follows:

$$\mathcal{L}^c_{RNA}(u, v) = \lambda_c \sum_{c=1}^{\mathcal{C}} \left(\frac{\mathbb{E}[h(X^u_c)]}{\mathbb{E}[h(X^v_c)]} - \frac{\mathbb{E}[h(X^v_c)]}{\mathbb{E}[h(X^u_c)]} \right)^2$$

3 where λ_c weights the loss, and $\mathbb{E}[h(X^m_c)]$ denotes the average norm of the features of modality m for samples of class c, with C the total number of classes. We note that in computing $\mathcal{L}^c_{RNA}$ for the target, the pseudo-labels are used to assign the target samples to classes.

The second extension of $\mathcal{L}_{RNA}$ addresses the problem of different norms in different domains by re-balancing the average and per-class norms of features in each modality across domains, so that the network can focus on features that are more transferable between domains [20]. To this end, we include the following term in the RNA formulation:

$$\mathcal{L}^{mod}_{RNA}(m_s, m_t) = \mathcal{L}^g_{RNA}(m_s, m_t) + \mathcal{L}^c_{RNA}(m_s, m_t)$$

where $m \in \{u, v\}$. Combining the three components we have previously defined, the extended RNA formulation in DG settings becomes:

$$\mathcal{L}_{RNA} = \mathcal{L}^g_{RNA}(u_s, v_s) + \mathcal{L}^c_{RNA}(u_s, v_s) \quad (2)$$

and the one for UDA setting is:

$$\begin{aligned} \mathcal{L}_{RNA} =& \mathcal{L}^g_{RNA}(u_s, v_s) + \mathcal{L}^g_{RNA}(u_t, v_t) + \\ & \mathcal{L}^c_{RNA}(u_s, v_s) + \mathcal{L}^c_{RNA}(u_t, v_t) + \\ & \mathcal{L}^{mod}_{RNA}(u_s, u_t) + \mathcal{L}^{mod}_{RNA}(v_s, v_t) \end{aligned} \quad (3)$$

The individual contribution of the three losses is exemplified in Fig. 3. $\mathcal{L}^g_{RNA}$ globally aligns the norms of modalities for each domain. $\mathcal{L}^c_{RNA}$ aligns the norms of modalities per class for each domain. $\mathcal{L}^{mod}_{RNA}$ aligns the norms between domains, separately for each modality. Taken together, the three losses act synergistically. In DG, $\mathcal{L}^c_{RNA}$ supports the work of $\mathcal{L}^g_{RNA}$, which in turn facilitates the alignment of norms per class to a common value. The addition of $\mathcal{L}^{mod}_{RNA}$ in UDA helps the other two components to ensure that the average and per-class norms of the different modalities are also aligned between source and target.

3.5 Extension to multiple modalities

The RNA objective in Eqs. 2 and 3 can be trivially extended to more than two modalities. In DG, the loss can be rewritten as:

$$\mathcal{L}_{RNA} = \mathcal{L}_{RNA}(\mathcal{S}) = \sum_{i=1}^{M} \sum_{j=i+1}^{M} \mathcal{L}_{RNA}(i_s, j_s) \quad (4)$$

where i and j span the M modalities. Similarly, the UDA loss becomes:

$$\mathcal{L}_{RNA} = \mathcal{L}_{RNA}(\mathcal{S}) + \mathcal{L}_{RNA}(\mathcal{T}) + \sum_{i=1}^{M} \mathcal{L}^{mod}_{RNA}(i_s, i_t)$$

where $\mathcal{L}_{RNA}(\mathcal{S})$ and $\mathcal{L}_{RNA}(\mathcal{T})$ are the loss in Eq. 4 for the source and target domains, respectively.

3.6 Learning objective in UDA

In addition to the loss defined in Eq. 3, to further improve the domain invariant properties of the features (and thus reduce the divergence between domains), we apply an adversarial domain alignment [62, 63]. We follow the recipe used in other recent UDA work [6, 46, 48, 64], and introduce a classifier that predicts whether features are from the source or the target. This classifier is directly connected to the feature extractors via a Gradient Reversal Layer (GRL) [62]. The domain classification loss $\mathcal{L}_d$ is then multiplied by a weight λ_d and added to the total loss.

The loss we have introduced so far (i.e., the combination of $\mathcal{L}_{RNA}$ and $\mathcal{L}_d$) aims to improve the informative and domain invariant properties of the embeddings of the different modalities. However, these two loss components affect the feature extractors F^m and are not back-propagated through the classifier, which therefore only sees the source data and thus has no way to benefit from the target data. The result is that during training, the classifier focuses only on how best to integrate the multi-modal features to improve accuracy in the source domain, and completely ignores the *classification uncertainty* on target.

One approach commonly used in UDA to improve class discrimination in the target domain is to use a mutual information criterion [65] applied to the target data that not only minimizes the prediction uncertainty, but also promotes a uniform distribution of samples between classes. This is achieved through an Information Maximization (IM) loss defined as the difference between the average entropy of the outputs and the entropy of the average output:

$$\mathcal{L}_{IM} = -\mathbb{E}_{x \in \mathcal{X}_T} \sum_{c=1}^{\mathcal{C}} p_c(x) \log p_c(x) + \sum_{c=1}^{\mathcal{C}} \bar{p}_c \log \bar{p}_c$$

where C is the total number of classes, p_c is the posterior probability for class c, and $\bar{p}_c$ is the mean output score for the current batch.

When we put all the pieces together, we train the model in the UDA setting to minimize the following loss:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{RNA} + \lambda_d \mathcal{L}_d + \lambda_{IM} \mathcal{L}_{IM}$$

where $\mathcal{L}_{RNA}$ is from Eq. 3 and λ_{IM} is the IM loss weight.

4 Experiments

In this section, we aim to verify the effectiveness of our proposed approach through an empirical evaluation on different multi-modal benchmarks corresponding to a variety of datasets and tasks. These range from action classification (on EPIC-Kitchens-55 [66], EPIC-Kitchens-100 [67], and UCF-HMDB [46]) to object recognition (on ROD [68]) and fatigue classification (on CogBeacon [69]).

In the analysis, the results are obtained and presented as follows. When a dataset includes different domains, we optimized the models using the average accuracy over all the domain splits reported in the respective experimental protocol. Results were obtained using the same set of hyperparameters for all splits. Therefore, in the following, we excluded from evaluation the methods for which it was obvious (either from the description or from the available source code) that the hyperparameters were optimized for each split.

The rest of the section is organized as follows. We begin by introducing the experimental benchmarks used in our work (Sec. 4.1). The ablation study is given in Sec. 4.2, and the results are presented from Sec. 4.3 to Sec. 4.7.

4.1 Datasets

EPIC-Kitchens-55 (EK55). This is a large-scale egocentric video benchmark recorded by 32 participants in their own kitchens while performing unscripted activities [66]. RGB, Audio and Flow data are available in the dataset. To validate our approach, we use the experimental protocol defined in [6]. According to this protocol, (i) we only use the three kitchens with the largest amount of annotated samples (hereafter referred to as D1, D2, and D3) and (ii) we consider only verb classification task and a subset of eight labels. The challenges lie not only in the large domain shift that exists between the different kitchens, but also in the unbalanced distribution of classes within and between domains.

EPIC-Kitchens-100 (EK100). EK100 [67] extends EK55 to 45 kitchens, with almost 100 hours of video and 89,977 annotated action segments. The UDA setting of EK100 is defined as two domains, *Source* (containing labeled training data from 16 participants collected in 2018) and

Target (i.e., unlabeled videos from the same 16 participants in the same kitchens but collected two years later). The segments are annotated with 97 nouns and 300 verbs corresponding to 3,369 unique action classes, largely unbalanced and characterised by a long-tailed distribution.

UCF-HMDB. The UCF-HMDB dataset [46] was published to study video domain adaptation in third-person action classification. The dataset consists of 3,209 videos from the original UCF101 [70] and HMDB51 [71] datasets, which define the source and target domains used in DG e UDA. The videos are annotated with 12 classes.

ROD. ROD [68] is an image-based dataset developed for object recognition tasks. ROD consists of 41,877 samples of 300 everyday objects grouped into 51 categories and captured by an RGB-D camera. ROD is coupled with SynROD [9], which contains photorealistic renderings from 3D models of the same categories as ROD, and N-ROD [72], which extends both datasets to event modality by introducing real event recordings obtained from ROD samples, as well as simulated events extracted from SynROD's synthetic images. Following the settings proposed in [9, 72], this dataset allows the exploration of domain shift between synthetic (source domain) and real data (target domain) in a multi-modal object classification task using RGB-Depth and RGB-Event.

CogBeacon. CogBeacon is a multi-modal dataset collected to analyze the effects of cognitive fatigue on human performance [69]. Volunteers completed three different computerized versions (V1, V2, and V3) of the Wisconsin Card Sorting Test, a test widely used in experimental and clinical psychology [73]. Experimental sessions are divided into rounds in which subjects can signal their cognitive fatigue (i.e., sample classes are "fatigue" and "no-fatigue", with a strong class imbalance towards *fatigue* in split V3). Two modalities are available: (i) EEG data, and (ii) user's movements and facial expressions (recorded by capturing 68 facial keypoints and the face bounding box).

4.2 Ablation studies

In this section, we present the ablation studies of our approach, all of which have been performed using EK100, as this is the largest and most diverse of all the benchmarks used in our work, thus increasing the statistical significance of these studies.

4.2.1 Experimental settings

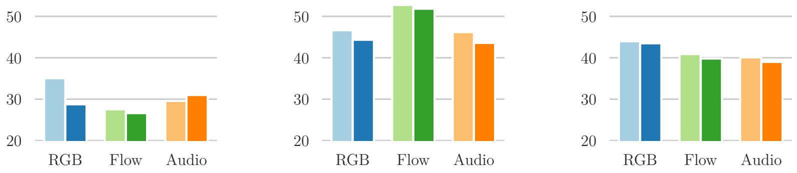
For the sake of clarity, we introduce here the details of the experimental settings used to obtain the results discussed in this Section and in Section 4.3.

Evaluation Protocol. We follow the experimental setup for UDA proposed in [67], where the fine-grained nature of the dataset annotations combined with the large domain and temporal shifts between the source and target domains make the adaptation task very challenging. All the experiments in this section (and in Section 4.3) use all three modalities (RGB, Audio, and Flow) available in the dataset. The setting includes a validation split, for which labels are available, and a non-annotated test split. The results of this work are reported on the former, although previous work has also demonstrated the effectiveness of RNA on test data as well [74, 75]. Performance is evaluated in terms of Top-1 and Top-5 accuracy of verb and noun predictions and on the combination of the two predictions (action).

Input. RGB, Flow and Audio are processed following [76] by uniformly sampling 25 frames and 1.28 seconds audio segments along the action. During both training and inference, five of these segments are selected for each modality and fed to the network.

Implementation Details. Frame-level features $f_m \in \mathbb{R}^{25\times1024}$ from each modality m are extracted using a TBN architecture [76] pretrained on Kinetics [77] and fine-tuned on the source domain, following the recipe from [67]. Our model is trained on pre-extracted features using this backbone. Five frame features for each segment are uniformly selected and fed to a linear layer, followed by a ReLU activation and dropout with probability 0.5. Frame features are temporally aggregated using a TRN [78] module to obtain action-level features $f'_m \in \mathbb{R}^{1024}$.[1] To account for the multi-task nature of this setting, we map the features into two components $f'_{m,v}$, $f'_{m,n} \in \mathbb{R}^{256}$ using a single linear layer,

[1] Up to this point, the implementation closely follows the official code provided for the EK100 UDA challenge [67].



(a) Source only

(b) $\mathcal{L}_{RNA}$ (DG)

(c) $\mathcal{L}_{RNA}$ (UDA)

Fig. 4 Verb feature norms across different modalities and settings (DG and UDA). Light (🟦 🟩 🟧) and dark colors (🟦 🟩 🟧) denote source and target validation domains, respectively. **(a)** In the "Source Only" setting, different modalities and domains result in unbalanced feature norms. **(b)** $\mathcal{L}_{RNA}$ in DG improves the alignment between different modalities, but leaves a gap between the source and target domains. **(c)** Finally, the contribution of $\mathcal{L}^{mod}$ in $\mathcal{L}_{RNA}$ reduces this gap in UDA, resulting in more consistent feature norms across different modalities and domains.

which we call *verb* and *noun features*. These are fed to two separate classifiers to obtain the modality logits for the verb ($y_{m,v}$) and the noun ($y_{m,n}$). Since this benchmark includes a single source and a single target domain, the network is trained for action recognition by applying cross-entropy loss to the sum of *per-modality* logits. We extend RNA to work in this multi-task context by applying the alignment losses separately to the verb and noun features, immediately before the final classifier. Applying the RNA losses to these features ensures that the alignment effect provided by RNA is as close as possible to the classifier, which is heavily influenced by the feature norm values. The network is trained for 30 epochs using a batch size of 128 samples and SGD optimizer with momentum 0.9 and weight decay 10^{-4}. The learning rate is initially set to 0.003 and decreased by a factor of 10 after epochs 10 and 20.

4.2.2 Effects of $\mathcal{L}_{RNA}$ on norm alignment

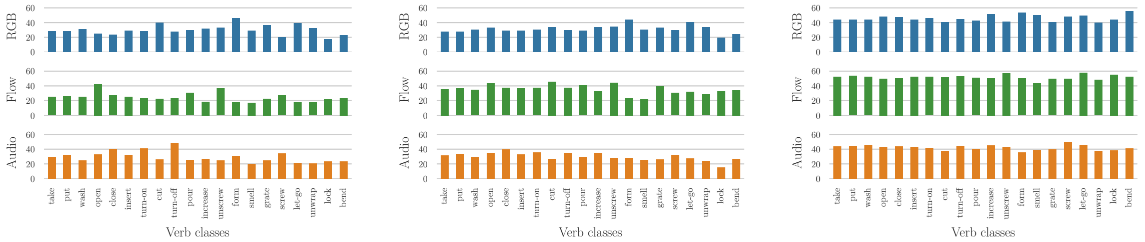
We begin by discussing the contribution of the components of the proposed $\mathcal{L}_{RNA}$ loss. Its goal is to mitigate domain shift issues by balancing the mean feature norms of the different modalities globally ($\mathcal{L}^{g}_{RNA}$), at the class level ($\mathcal{L}^{c}_{RNA}$), and across domains ($\mathcal{L}^{mod}_{RNA}$). In the following, we present the results of experiments in which these components are introduced incrementally.

Global alignment: a qualitative analysis. In Fig. 4 we report the mean feature norms for each modality. For simplicity, we will base our discussion on the verb feature norms, since the same observations apply to nouns. In particular, in Fig. 4 we show how the average norms of verb features for different modalities change on DG and UDA with the contribution of $\mathcal{L}_{RNA}$.

A preliminary qualitative analysis of the data presented in Fig. 4 shows that $\mathcal{L}_{RNA}$ in DG (Fig. 4b) leads to a better alignment of the average feature norms of the different modalities and to an overall increase of their values with respect to the "Source Only" (Fig. 4a). Recall that the norm formulation in Eq. 2 attempts to solve the alignment task at the batch level, and thus does not guarantee an exact alignment of all average norms. In Fig. 4b, we can also observe the increase in Flow norm in DG compared to "Source Only" (Fig. 4a). Previous studies have shown that Flow is the modality least affected by domain shift in egocentric action recognition [6], potentially allowing for greater generalization. This could explain why, in DG, the network pays more attention to this modality.

In addition, the availability of target data in UDA enables $\mathcal{L}_{RNA}$ to improve the balance between the norms of the different modalities, so that the model can better use the contributions of each modality to make its final decisions. This improved mutual contribution between modalities (reflected in the increased accuracy reported in Table 2) may explain the (relatively) lower norm of Flow in UDA, which is balanced by increased norms of (i.e., attention to) the other two modalities (RGB and Audio).

Global alignment: a quantitative analysis. To facilitate the assessment of the balancing effect of $\mathcal{L}_{RNA}$ between "Source Only", DG and UDA norms, we also introduce a quantitative metric. We use the *coefficient of variation* (CV) as a measure of the norm imbalance, with lower CVs indicating more balanced sets of values. CV is



(a) Source Only

(b) $\mathcal{L}^g_{RNA}$

(c) $\mathcal{L}^g_{RNA} + \mathcal{L}^c_{RNA}$

Fig. 5 Feature norms of the top 10 most and least common classes from the target validation split of EPIC-Kitchens-100. While $\mathcal{L}^g_{RNA}$ improves the alignment of different modalities, there is still an imbalance between classes. The addition of the per-class variant of RNA greatly improves this alignment, resulting in more uniform feature norms across different classes.

Method	**CV**$_S$	**CV**$_T$	**CV**$_{S+T}$
Source Only	0.126	0.076	0.101
$\mathcal{L}^g_{RNA}$	0.089 (+29.3%)	0.121 (-59.8%)	0.103 (-2.1%)
$\mathcal{L}^g_{RNA} + \mathcal{L}^c_{RNA}$ (DG)	0.075 (+40.4%)	0.098 (-28.7%)	0.081 (+20.1%)
$\mathcal{L}_{RNA}$ (UDA)	0.049 (+61.0%)	0.059 (+22.7%)	0.049 (+50.7%)

Table 1 Coefficient of variation for DG and UDA feature norms. CV$_S$, CV$_T$ and CV$_{S+T}$ are the CVs of the source, target and combined domain(s) respectively. For clarity, we also report the percentages of improvement with respect to the "Source Only" experiment.

defined as follows:

$$CV = \frac{\sigma}{\mu}$$

where σ is the standard deviation and μ is the mean of the observed norm values. The CV values obtained are summarized in Table 1, where, for better clarity, we also report the percentage of improvement (%) with respect to the CV values of the "Source Only".

As for the average feature norms in DG (Fig. 4b), we have a 40.4% decrease in CV compared to the "Source Only". It is interesting to note that the application of $\mathcal{L}^g_{RNA}$ alone only contributes to a 29.3% reduction of CV, highlighting the (positive) combined effect of $\mathcal{L}^g_{RNA}$ and $\mathcal{L}^c_{RNA}$. For the target domain in DG, we can observe that the imbalance between modalities increases (instead of decreasing) by 28.7%, which highlights the need for an alignment loss that works not only between modalities but also between domains.

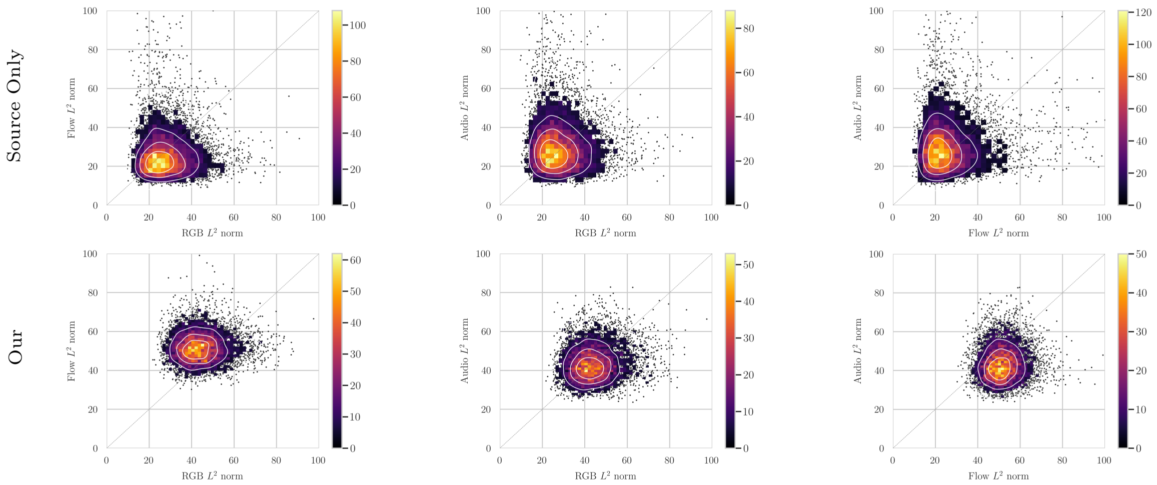
In UDA, the ability to use the target data contributes to a larger reduction in CV over the "Source Only" on both source (by 61.0%) and target domains (22.7%). When we consider the total imbalance (i.e., we calculate CV considering all source and target values together), CV shows an improvement of 20.1% in DG and of 50.7% in UDA. These values are reflected in progressively greater accuracy in the DG and UDA settings compared to the "Source Only" settings (Table 2).

Class alignment. For assessing the contribution of $\mathcal{L}^c_{RNA}$, we show in Fig. 5 the evolution of the verb norms of the ten most frequent and the least frequent classes in the DG settings. In the "Source Only" (Fig. 5a) the per-class mean features norms are largely unbalanced. While the exclusive use of $\mathcal{L}^g_{RNA}$ contributes to a better global balance of the modality norms, it has a small effect on the balancing of the norms per-class (Fig. 5b). On the contrary, when $\mathcal{L}^c_{RNA}$ is also minimized, we can observe a significant improvement of their alignment (Fig. 5c).

These qualitative observations are also reflected in the CV metric computed on the class norms. Indeed, the use of $\mathcal{L}^g_{RNA}$ leads to a minor improvement in "Source Only" CV (37.8% and 19.5%, respectively, for source and target features) compared to that obtained by the combination of $\mathcal{L}^g_{RNA}$ and $\mathcal{L}^c_{RNA}$ (62.5% and 49.1%).

Overall effect on feature norms. To give further insight into the impact of $\mathcal{L}_{RNA}$, we show in Fig. 6 a scatter plot of the validation set in DG. This diagram is obtained by plotting the RGB, Flow and Audio feature norms of each sample in a 3D space whose axes are the norms of the three modalities. To make the plot easier to read, rather than using a single 3D representation, we present it as three separate sections along the coordinate planes defined by the feature pairs. The goal of these visualizations is to illustrate the changes in the shape of the resulting manifold.

It can be seen that the "Source Only" features are widely distributed and correspond to a



(a) Flow v. RGB (b) Audio v. RGB (c) Audio v. Flow

Fig. 6 Comparison of the feature norms before (top) and after (bottom) application of $\mathcal{L}^{g}_{RNA}$ and $\mathcal{L}^{c}_{RNA}$. The dots represent the samples in the validation dataset. The color bar on the right represents increasing density values. The original features, i.e. "Source Only", show a wide range of values and an irregular shape, reflecting the misalignment between the features norms of the two modalities. The RNA loss re-balances the two, as evidenced by the more globular distribution while also shifting the average norms towards higher values.

Method	Verb@1	Noun@1	Action@1	Δ Acc.
Source only	46.79	26.79	18.29	-
DG				
$\mathcal{L}^{g}_{RNA}$	49.53	27.50	18.91	1.36
$\mathcal{L}^{c}_{RNA}$	50.51	27.75	19.44	1.94
$\mathcal{L}^{g}_{RNA} + \mathcal{L}^{c}_{RNA}$	50.75	27.92	19.81	2.20
UDA				
$\mathcal{L}^{g}_{RNA}$	49.98	27.79	19.44	1.78
$\mathcal{L}^{g}_{RNA} + \mathcal{L}^{c}_{RNA}$	50.46	28.49	19.77	2.28
$\mathcal{L}^{g}_{RNA} + \mathcal{L}^{c}_{RNA} + \mathcal{L}^{mod}_{RNA}$	49.94	**29.48**	19.87	2.48
$\mathcal{L}_{RNA} + \mathcal{L}_{d}$	50.59	29.38	20.04	2.71
$\mathcal{L}_{RNA} + \mathcal{L}_{IM}$	**51.04**	28.86	19.97	2.67
$\mathcal{L}_{RNA} + \mathcal{L}_{d} + \mathcal{L}_{IM}$	50.82	29.19	**20.05**	2.73

Table 2 Ablation on different loss components. Δ Acc. is the average accuracy improvement for the verb, noun and action metrics. Best in **bold** and the runner-up underlined.

manifold with a largely irregular shape. This is due to misalignment between the feature norms of the different modalities. When the $\mathcal{L}_{RNA}$ loss is applied, the manifold becomes more spherical and compact, reflecting the improvement in the alignment of the modality norms. It is also possible to note an increase in the average feature norm values that moves the manifold towards the upper right region of the 2D dimensional plots.

4.2.3 Effect of loss components

Table 2 details the contribution of the different loss components to the final performance in both DG and UDA settings. For better evaluation, we also show the average improvement in Top-1 accuracy of verb, noun, and action with respect to "Source Only" (Δ Acc). The combination of global and class components in DG ($\mathcal{L}^{g}_{RNA} + \mathcal{L}^{c}_{RNA}$, Δ Acc. = 2.20) improves accuracy over $\mathcal{L}^{g}_{RNA}$ and $\mathcal{L}^{c}_{RNA}$ alone (1.36 and 1.94, respectively), showing

Method	Verb@1	Noun@1	Action@1	Δ Acc.
RGB + Flow				
Source Only	44.80	25.35	16.33	-
Our (DG)	45.95	26.65	16.94	1.02
Our (UDA)	47.64	26.49	16.91	1.52
RGB + Audio				
Source Only	39.91	24.18	14.84	-
Our (DG)	42.04	25.54	15.67	1.44
Our (UDA)	42.26	26.45	15.98	1.92
Flow + Audio				
Source Only	45.11	21.98	15.37	-
Our (DG)	48.87	23.44	16.49	2.12
Our (UDA)	48.42	23.51	16.71	2.06
RGB + Flow + Audio				
Source Only	46.79	26.79	18.29	-
Our (DG)	50.75	27.92	19.81	2.20
Our (UDA)	50.82	29.19	20.05	2.73

Table 3 Top-1 classification accuracies (%) on modality pairs on EPIC-Kitchens-100 [67]. Δ Acc. is the average accuracy improvement for the verb, noun and action metrics.

that the combination of the two components effectively reduces domain shift. Indeed, while L^g_{NRA} L^g_{NRA} aligns modalities globally, possibly penalizing minority classes in unbalanced distributions, L^c_{NRA} enforces alignment for each class individually. The ability to use target data in UDA boost the accuracy improvement to 1.78 for $\mathcal{L}^g_{RNA}$ and 2.28 for $\mathcal{L}^g_{RNA} + \mathcal{L}^c_{RNA}$), with $\mathcal{L}^{mod}_{RNA}$ further contributing to reach an average improvement of 2.48.

As explained in Sec. 3.6, the learning objective in the UDA setting also benefits from two other losses, namely the adversarial domain loss $\mathcal{L}_d$, which aims to improve the transferability of features across domains, and the Information Maximization loss $\mathcal{L}_{IM}$, which aims to minimize the classification uncertainty between target classes. $\mathcal{L}_d$ provides a stronger improvement in this particular case (2.71), while $\mathcal{L}_{IM}$ has a minimal effect on the overall accuracy. However, we note that the mutual contribution of the latter two terms ($\mathcal{L}_d$ and $\mathcal{L}_{IM}$) also depends on the task and benchmark considered, as other experiments show more pronounced benefits for $\mathcal{L}_{IM}$.

4.2.4 Multi-modal adaptation capabilities

Another interesting question is whether the proposed method allows effective integration of multiple modalities in the final decision and whether the use of multiple modalities also helps to improve the domain adaptation capabilities of the model.

Table 3 summarizes the results obtained comparing experiments with modality pairs and with all three modalities. It shows that the latter not only outperforms all other modality pairs in terms of results, but also shows better generalization properties, showing an improved delta compared to its "Source only" (2.73) compared to 2.06, the best two-modality improvement obtained with Flow + Audio. These results suggest that our method is effective in combining the different modalities to improve the overall accuracy and the generalizability of the features obtained.

4.2.5 Modality drop

In Table 4, we present an experiment to investigate the impact of modality imbalance during training. In particular, we investigate the scenario in which a modality is "unexpectedly" lost at inference time, without a training strategy accounting for this possibility. This scenario, also presented in [79], is relevant because there may be constraints at inference time, such as power, computational or privacy constraints, or an anomaly of an input device that prevent the use of all modalities.

Method	Verb@1	Noun@1	Action@1	Δ Acc.
No Audio @ Test				
Source only	41.61	21.91	13.07	-
DG	44.03	24.44	14.89	2.26
UDA ($\mathcal{L}_{RNA}$)	44.08	24.77	15.25	2.50
No Flow @ Test				
Source only	30.58	20.33	10.63	-
DG	36.88	22.82	12.89	3.69
UDA ($\mathcal{L}_{RNA}$)	36.67	21.83	12.46	3.14
No RGB @ Test				
Source only	37.69	17.99	12.41	-
DG	46.70	18.92	13.53	3.69
UDA ($\mathcal{L}_{RNA}$)	46.51	19.37	13.55	3.78

Table 4 Modality drop. All configurations are trained on all input modalities. At inference time, we simulate the loss of a modality, resulting in large performance drops that RNA helps mitigate.

The basic idea of our approach is to help the model learn equally from the different modalities by integrating their contribution. While it is clear that the unexpected loss leads to a drop in accuracy, we can also expect that the effect of RNA is to make the model more robust to such a modality drop than the "Source Only" model, since the latter is less able to exploit the synergies between modalities and, thus, more vulnerable to dominant modalities. This expectation is confirmed by the results in Table 4, which are consistent with the observations of [79], and show different but consistent effects on "Source Only" when different modalities are dropped at test time (i.e., large accuracy drops compared to the results in Table 2). At the same time, these results show that the balancing effect of RNA can potentially help the model reduce the impact of the lost modality, as it can take advantage of a better mutual contribution from the remaining ones.

4.3 Experiments on EK100

Unlike the following benchmarks, where we describe the experimental protocol, inputs, implementation details, and baselines used to evaluate the results, in this section we present only the baselines for EK100, as the previous elements were introduced in Sec. 4.2.1.

Baselines. We compare our method with MM-SADA [6], TA^{3}N [46], and CIA [54]. As for MM-SADA, the original approach works only with RGB and Flow modalities. Therefore, to integrate the Audio modality, we use two separate branches, one for RGB-Flow and the other for RGB-Audio modalities (as in [16]). The adversarial branch is applied individually to each modality.

Results. Results are given as Top-1 and Top-5 accuracy for verb, noun, and action. For each baseline, we also report the relative "Source Only", average improvement in terms of Top-1 accuracy. For the DG setting, we compare our approach to two alternative methods. The first is MM-SADA (SS), a modified version of MM-SADA, which applies only the original self-supervised alignment task to the source domain modalities and does not consider the adversarial alignment component of the method (which requires target data). The second approach is Gradient Blending (GB), which attempts to find an optimal mixture of modalities according to their overfitting behavior. Such a mixture is achieved by combining a cross-entropy loss for each modality and a loss for their fusion with appropriate weights[2].

Analyzing the accuracy across different labels, we observe that GB performs best, while our approach ranks as the runner-up and MM-SADA (SS) lags slightly behind. However, when considering the improvements relative to the "Source Only" baseline, our method shows higher delta accuracy compared to GB. This result seems to indicate that our method makes a more significant contribution to reducing the domain shift. We also find the approach proposed in [11] interesting as it shares similarities with our method in terms of improving the balance between modalities for better classification accuracy. To investigate this further, we perform additional experiments by applying our method to the "Source Only" results obtained from Gradient Blending, i.e., using multiple classification losses but without reweighting them. These additional experiments are indicated with a † symbol. The results shown in Table 3 are promising. Our method achieves the best action accuracy and is competitive with GB (and also comparable with CIA, the state-of-the-art in UDA). It is important to note that our standard solution addresses the alignment problem with an adaptive approach that, unlike GB, is independent of the model and dataset used and requires only two hyperparameters: λ_g and λ_c.

In the UDA experiments, we observe that although our method ranks second in terms of action accuracy, it has better delta accuracy improvements compared to all other competitors. Furthermore, the results on other evaluation metrics are comparable to those of the other proposed baselines. It is noteworthy that a significant portion of the improvements can be observed in the DG phase, where the target domain is not accessed. This observation highlights the generalization advantage of RNA in coping with domain shifts.

4.4 Experiments on EK55

Evaluation Protocol. We adopt the experimental protocol of [6] and evaluate performance in a single-source setting ($D_i \rightarrow D_j$) on the three domains described in Sec. 4.1. Despite the small size of this setting compared to EK100, it remains a highly valued and challenging benchmark in the field of egocentric action recognition due to the large domain shift between these domains and the unbalanced label distribution. In the experiments, we restrict our analysis to the RGB+Flow and RGB+Audio modality combinations, which are the ones recent work in the literature focus on.

Baselines. We compare our results with several state-of-the-art UDA methods. The first group (GRL [28], MMD [41], AdaBN [81], and MCD [40]) includes approaches originally developed as image-based methods and later adapted to work with video inputs. The second group includes more recent methods such as MM-SADA [6], the contrastive-based methods proposed in [23] and [22] (STCDA), and the recently published CIA [54]. In our comparison, we use the results reported in the original paper for each baseline.

Input. As for the input, different sampling strategies are used to allow a fair comparison with the existing baselines. When using *dense sampling*, a clip of 16 consecutive frames is randomly

[2] The original version of GB uses only RGB and Audio. The optimal weights for combining losses were taken from [80], and the weight for the missing component, i.e., Flow, was tuned appropriately for this work

Methods	Network	Verb@1	Noun@1	Action@1	Verb@5	Noun@5	Action@5	Δ Acc.
DG								
Source Only	TBN-TRN	47.14	27.35	18.99	75.27	49.36	41.82	-
MM-SADA (SS) [6]	TBN-TRN	47.76	27.93	19.15	77.07	49.77	42.90	0.45
Source Only	TBN-TRN	50.27	29.04	19.96	81.74	52.14	46.74	-
GB [11]	TBN-TRN	50.18	**29.60**	20.26	**81.82**	52.57	**46.86**	0.26
Source Only	TBN-TRN	46.79	26.79	18.29	75.39	48.44	41.36	-
Our (DG)	TBN-TRN	**50.75**	27.92	19.81	80.64	51.37	45.33	2.20
Source Only†	TBN-TRN	49.81	28.55	19.77	81.10	51.90	46.22	-
Our† (DG)	TBN-TRN	50.20	29.31	**20.30**	81.58	**52.68**	46.76	0.56
UDA								
Source Only	TBN-TRN	46.70	27.78	19.20	75.42	48.27	42.12	-
TA3N [46]	TBN-TRN	48.44	28.87	19.61	75.95	50.12	43.36	1.08
Source Only	TBN-TRN	47.14	27.35	18.99	75.27	49.36	41.82	-
MM-SADA [6]	TBN-TRN	48.44	28.26	19.25	77.56	50.59	43.41	0.82
Source Only	TBN-TRN	47.69	28.48	19.61	-	-	-	-
CIA [54]	TBN-TRN	48.34	**29.50**	**20.30**	-	-	-	0.79
Source Only	TBN-TRN	46.79	26.79	18.29	75.39	48.44	41.36	-
Our (UDA)	TBN-TRN	**50.82**	29.19	20.05	**80.89**	**52.18**	**46.04**	2.73

Table 5 Classification accuracies (%) on EPIC-Kitchens-100 [67]. Results are reported in terms of Top-1 and Top-5 classification accuracy for the noun, verb and action metrics. Δ Acc. is the average Top-1 accuracy improvement. †These experiments are trained using the cross entropy loss on both the fused logits as well as on the *per-modality* logits. Best in **bold**, runner-up underlined.

sampled from the video. When using *uniform sampling*, 16 frames evenly distributed over the video are sampled. At test time, the same sampling strategy is used as in training, except that five clips are fed into the network instead of one, as suggested in [82], and the predictions are averaged. Following the experimental setting from [6], during training, random clipping, scale shifts, and horizontal flipping are used for data augmentation, while in testing, only central cropping is applied. As for the aural information, we follow [76] 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 and sampled at 24kHz. Then, the Short-Time Fourier Transform (STFT) is calculated with a window length of 10ms, a skip size of 5ms, and 256 frequency bands. For the Flow input, we use the same sampling strategy as for RGB.

Implementation Details. Both the RGB and Flow streams use an I3D model [83] pre-trained on Kinetics [77], following the experimental setting from [6]. Following [76], the audio feature extractor uses the BN-Inception model [84] pre-trained on ImageNet [85]. The feature extraction backbones are trained end-to-end. For each modality m, features have shape $f_m \in \mathbb{R}^{1024}$. Logits are computed separately for each modality using a linear layer and summed. We train the network for 5000 iterations using the SGD optimizer using momentum 0.9 and weight decay 10^{-7}. The learning rate for RGB and Flow is set to 0.001 and reduced to 2×10^{-4} at step 3000, while for Audio the learning rate is set to 0.001 and decremented by a factor of 10 at steps $\{1000, 2000, 3000\}$. The batch size is set to 128.

Results. We begin by discussing the UDA results, which are summarized in Table 6. Given the relevance of sampling strategies in the video context, especially for the RGB+Flow combination [86], we divide the results into different sections based on the sampling used for each modality: dense (D) or uniform (U). Most baselines use dense sampling (D-D), while CIA is the only method that uses uniform sampling (U-U) for both modalities. In both cases, we compare the baselines to our UDA method using the same sampling strategy.

The results show that CIA with uniform sampling outperforms the dense sampling-based methods. This observation confirms the findings in [86], which emphasizes that uniform sampling usually allows the network to learn more information. We also observe that our UDA approach achieves state-of-the-art results for both dense and uniform samplings.

Method	Sampling	D1→D2	D1→D3	D2→D1	D2→D3	D3→D1	D3→D2	Mean
				RGB + Flow				
Source Only	D-D	42.00	41.20	42.50	46.50	44.30	56.30	45.47
GRL [28]	D-D	50.20	44.70	46.90	50.80	50.20	53.60	49.40
MMD [41]	D-D	46.60	39.20	43.10	48.50	48.30	55.20	46.82
AdaBN [81]	D-D	47.00	40.30	44.60	48.80	47.80	54.70	47.20
MCD [40]	D-D	46.50	43.50	42.10	51.00	47.90	52.70	47.28
DAAA [48]	D-D	50.00	43.50	46.50	51.50	51.00	53.70	49.37
MM-SADA [6]	D-D	49.50	44.10	48.20	52.70	50.90	56.10	50.25
Kim et al. [23]	D-D	50.30	46.30	49.50	52.00	51.50	56.30	50.98
STCDA [22]	D-D	52.00	45.50	49.00	52.50	52.60	55.60	51.20
Our (UDA)	D-D	50.84	47.14	48.86	54.38	50.6	58.43	**51.71**
Source Only	U-U	43.20	42.50	43.0	48.0	43.0	55.50	45.90
CIA [54]	U-U	52.50	47.80	49.80	53.20	52.20	57.60	52.18
Our (UDA)	U-U	52.84	47.49	54.41	54.11	55.53	61.64	**54.34**
Source Only	D-U	54.25	50.72	54.87	56.41	51.65	61.27	54.86
Our (DG)	D-U	56.00	50.39	56.25	56.37	56.73	61.63	56.23
Our (UDA)	D-U	57.33	52.84	57.19	56.78	57.27	62.03	**57.24**
				RGB + Audio				
Source Only	D-D	39.03	39.17	35.27	47.52	40.255	49.98	41.87
GRL [28]	D-D	41.02	43.04	39.36	49.25	38.77	50.56	43.67
MMD [41]	D-D	42.40	43.84	40.87	48.13	41.46	50.03	44.46
AdaBN [81]	D-D	36.64	42.57	33.97	46.63	40.51	51.2	41.92
MM-SADA [6]	D-D	48.90	46.66	39.51	50.89	45.42	55.14	47.75
Our (DG)	D-D	42.55	41.77	42.73	51.09	42.63	54.24	46.21
Our (UDA)	D-D	46.65	47.22	46.18	52.30	44.04	56.18	**48.76**

Table 6 Classification accuracies (%) on EPIC-Kitchens-55 [66], using the evaluation protocol from [6], divided by modalities. Results are grouped by the sampling strategy used for a fair comparison. Best in **bold**, runner-up underlined.

To further confirm the importance of sampling, we conduct experiments with a mixed sampling strategy (i.e., D for RGB and U for Flow, Table 6). Since none of the baselines use this sampling, we only present our results for the "Source Only", DG, and UDA methods. We note that the "Source Only" method already achieves remarkable results (up to 3% better than our method with uniform sampling), which are further improved in both DG and UDA (despite the smaller difference with the "Source Only" compared to other samplings). One possible explanation for the improved performance with mixed sampling is that it allows for better exploitation of the distinct properties of the two modalities. Dense sampling facilitates a more accurate characterization of static appearance information (RGB) over a short temporal range, while uniform sampling enables the use of a wider temporal range to capture the dynamic information conveyed by Flow.

When combining RGB and Audio modalities, our UDA approach consistently achieves the best results (7% improvement over "Source Only" and 1% improvement over the state-of-the-art method). This result confirms the potential of our method, even when dealing with the fusion of heterogeneous modalities.

Finally, we discuss the results we obtained in DG for both modality combinations. For RGB+Flow, we report the results obtained with the mixed sampling strategy (D-U), i.e., the sampling that yields the best performance. In this setting, our method improves the "Source Only" by up to 2% and 5% for RGB+Flow and RGB+Audio, respectively. Furthermore, the performance obtained in the DG setting is comparable to that of the UDA setting, with a deviation of -1.01% and -2.55% for RGB+Flow and RGB+Audio, respectively. Although no other DG methods are available for comparison in this context, these results show that the DG setting can compete with several existing UDA methods that benefit from target data during training.

4.5 Experiments on UCF-HMDB

Evaluation Protocol. We follow the same experimental setting proposed in [78], which includes the U → H and H → U shifts in a multi-modal setting that includes the RGB and Flow modalities available with this dataset.

Method	U→H	H→U	Mean
Source Only	82.8	90.7	86.7
MM-SADA [6]	84.2	91.1	87.6
Source Only [22]	82.8	89.8	86.3
STCDA [22]	83.1	92.1	87.6
Source Only [23]	82.8	90.7	86.7
Kim et al. [23]	84.7	92.8	88.7
Source Only [54]	86.1	92.5	89.3
CIA [54]	88.3	94.1	91.2
Source Only (conc) [54]	85.8	93.5	89.5
CIA (conc) [54]	90.6	94.2	**92.4**
Source Only	83.6	94.1	88.9
Our (DG)	83.3	94.9	89.1
Our (UDA)	86.4	94.3	90.4

Table 7 Classification accuracies (%) on UCF-HMDB on RGB+Flow combination. Best in **bold**, runner-up underlined.

Input. For both RGB and Flow, the training input consists of 16 consecutive frames with resolution 224 x 224 pixels. In testing, we use five clips uniformly sampled across the video and average the predictions. We use the same data augmentations as described in Sec. 4.4 for EK55.

Baselines. We compare our approach with various multi-modal UDA approaches (MM-SADA [6], STCDA [22], the method of Kim et al. [23] and CIA [54]). To allow a fair comparison, all multi-modal results are based on the same backbones and the same pre-training.

Implementation Details. The backbone for both RGB and Flow is an I3D pre-trained on Kinetics [77]. The learning rate is set to 0.01 and we train the model for 20 epochs with batch size of 32. We use SGD as the optimizer with a momentum of 0.9 and a weight decay of 10^{-7}.

Results. In Table 7, we present the classification accuracy of our method and several baselines. To ensure a fair comparison, we report the results of the "Source Only" model from the original paper for all baselines.

In absolute terms, our approach achieves very competitive performance under the UDA setting and is the second best of all methods in terms of accuracy, outperforming all baselines except CIA. However, the better "Source Only" result of CIA, which was difficult to reproduce in our experiments, should be emphasized. This result could be attributed to its particular architectural design choices and the integration of spatial consensus between RGB and Flow modalities. However, it should be noted that the method proposed by CIA cannot be easily extended to other modalities, as shown by their work on integrating the Audio modality in EK100 [54]. In contrast, our approach provides a more versatile and adaptable solution that can be applied to different modalities without significant architectural changes.

Furthermore, our approach shows remarkable domain shift reduction capabilities. When we compare the performance gains of our method with other baselines, we observe that our approach achieves improvements over the "Source Only" baseline that are comparable to those obtained by other methods. For example, MM-SADA, STCDA, Kim et al. [23], and CIA show gains of 0.9%, 1.3%, 2%, and 1.9%, respectively, while our approach achieves a gain of 1.5%, with a maximum improvement of up to 3% in the U → H shift. This highlights the effectiveness of our method in adapting to target domains and mitigating the negative effects of domain shift.

4.6 Experiments on ROD

Evaluation Protocol. We follow the experimental protocol in [9] for RGB-depth modalities, and the one in [72] for RGB-event. The studied shift is a synthetic-to-real domain shift, with synthetic source data and real target data (SynROD → ROD). RGB and depth modality in the synthetic domain are rendered, while events in the synthetic domain are simulated using ESIM [87].

Baselines. We compare our results with standard image-based UDA methods, namely GRL [28], MMD [41], SAFN [20] and Entropy [88], which we extend to operate on multiple modalities. We also compare with Relative Rotation [9], a method specifically designed to operate on multiple modalities. It consists in a self-supervised task asking the network to predict the relative rotation between two modalities of the same input sample, e.g., an RGB and a depth image.

Input. Event representations, depth images and RGB images are pre-processed and augmented during training following the procedure in [9]. Depth images are colorized with surface normal encoding, as in [89]. Input images are normalized with the same mean and variance used for the ImageNet pre-training, while we kept event representations un-normalized as this provided better performance. We use voxelgrid representation for events with 9 bins as in [72].

Implementation Details. All backbones are implemented using ResNet-18 [90], pre-trained on ImageNet [85]. All the parameters of the network, including the pre-trained parameters, are updated during training, as in [9]. We train all network configurations using SGD as optimizer, batch size 64 and weight decay 0.003.

Results. Table 8 provides a comparison of our method with different baselines for the SynROD→ROD adaptation task using RGB, depth, and event modalities.

In UDA setting, our method shows remarkable performance gains over all baselines for both RGB+Depth and RGB+Event combinations. For RGB+Depth, our approach achieves improvements of up to 20% over the baselines. these results deomonstrate the capability of our method in reducing the domain shift and improving classification accuracy when adapting from the synthetic domain (SynROD) to the real-world domain (ROD). Our method also achieves improvements of up to 10% over the baseline values for the RGB+Event combination. This is a further evidence of the effectiveness of our method in handling the domain shift and improving classification accuracy even in the presence of event data.

However, for DG, the performance gains are relatively smaller compared to the *Source Only* model for both modality combinations. This can be attributed to the inherent challenges of the synthetic-to-real setting, where there is a significant gap between the feature distributions of the source (SynROD) and target (ROD) domains. The unavailability of target data during training limits the generalization capabilities of the model, resulting in modest improvements over *Source Only*. In contrast, in UDA setting, our method achieves better performance by effectively using the target domain information to bridge the domain gap.

Thus, we can conclude that the success of our method in mitigating the shift from synthetic to real domains highlights its potential for various fields and applications, such as robotics, autonomous driving, and augmented reality, where synthetic training data are largely used.

4.7 Experiments on CogBeacon

Evaluation Protocol. We follow the experimental protocol in the supplemental of [15], evaluating the performance in the single-source setting ($V_i \rightarrow V_j$) using three different domains (V1, V2, and V3), for a total of six splits.

	RGB + D	RGB + E
Source Only	47.70	49.19
GRL [28]	59.51	55.11
MMD [41]	62.57	62.39
SAFN [20]	62.40	66.87
Entropy [88]	63.12	66.23
Relative Rotation [9]	66.68	66.68
Our (DG)	50.06	50.61
Our (UDA)	**82.36**	**78.52**

Table 8 SynROD→ROD accuracy (%) results. Best in **bold**, runner-up underlined.

Baselines. We compare our results with those in [15] (in particular with DLMM, the Differentiated Learning framework proposed in [15]) and with those obtained in our experiments with different UDA methods, namely SAFN [20], GRL [28], MMD [41], and MM-SADA [6]. These two lists of results are presented separately in Table 9 because the number of samples does not match that used in [15] (i.e., we have 2,240, 2,432, and 2,300 for domains V1, V2, and V3, respectively).

Input. The EEG signals are characterized using a total of 24 temporal and spectral features (see [69] for details). The face data are represented as a vector combining the average values of the face data and their standard deviation, yielding a total of 280 values.

Implementation details. Both backbones are implemented by three 1D convolutional blocks with kernel size three and stride one, followed by a MaxPool layer and ReLU as the activation function. The output channels are 16, 32 and 64 for EEG signals and 8, 16 and 32 for the face keypoint model. The latter ends with a fully connected layer with an output of 64 to match the output of the EEG backbone. Model weights are randomly initialized. Predictions for each modality are computed with a single FC layer followed by a LogSoftmax. We train the model for 90 epochs using Adam as the optimizer. In all experiments, the learning rate was set to $1e{-}3$ and decremented by a factor of 10 after 70 epochs.

Results. In Table 9, we present the classification accuracy of our method and several baselines for the CogBeacon dataset.

Results from [15]		Ours	
Source Only	63.64	Source Only	61.80
MDANN [51]	66.83	SAFN [20]	64.01
MCD [40]	66.75	GRL [28]	64.24
CBST [91]	67.71	MM-SADA [6]	65.40
MM-SADA [6]	67.75	MMD [41]	65.58
DLMM [15]	**70.47**	Our (DG)	62.64
		Our (UDA)	**68.63**

Table 9 CogBeacon accuracy (%) results. Best in **bold**, runner-up underlined.

When we compare our approach to the "Source Only" baseline, we observe a significant improvement in the UDA setting. Our UDA accuracy of 68.63% significantly outperforms the "Source Only"'s accuracy of 61.80%. On the other hand, the improvement in the DG setting is not so significant, with an accuracy of 62.64%. However, even in this setting, we observe a potential for softening the domain shift and achieve better performance.

It is worth noting that, on our settings, our method outperforms the other UDA methods used for comparison. Among the UDA baselines, MMD achieves the highest accuracy of 65.58%, which is 3.05% less than our method. We are also competitive with more complex domain matching approaches such as CBST, which involve the generation of pseudo-labels or the use of confidence-based selection strategies [91].

However, compared to the results reported in [15], our method falls behind DLMM, which achieves 70.47% accuracy, significantly better than our results. Nevertheless, a more detailed analysis reveals interesting insights when considering the improvements over the "Source Only" baseline. DLMM achieves an improvement of 6.83%, while our method has an improvement of 7.03%, which can be considered equivalent. That said, it is worth mentioning that our approach is characterized by its simplicity compared to DLMM. DLMM requires multiple training stages and uses a more complex curriculum learning approach with teacher/student models for different modalities. In contrast, our method requires is lighter and simpler to train, making it a more practical option for real-world applications.

4.8 Limitations

The proposed approach provides interesting performance in many cases, as shown by our experiments with a variety of tasks and scenarios. While the simplicity of the method is certainly a strength, it may be viewed as less effective when compared to methods developed and tuned for a specific task and benchmark. However, we believe that this limitation does not undermine the overall effectiveness of the proposed approach, as it provides a viable alternative for addressing various tasks without requiring significant computational resources or architectural changes.

Another limitation we observed arises from the fact that in many real-world cases the data distributions are strongly unbalanced, leading to lower precision for the tail classes [92]. The literature shows how this imbalance translates into unbalanced norms of classification weights per class [93, 94] as well as unbalanced norms of features per class [95, 96]. In developing our method, we expected that balancing the norms per class could have a positive effect also in rebalancing the weights of the classifier for the tail classes. However, our experimental results show that this effect is not present. This opens up possibilities for future developments to incorporate this objective into RNA as an additional component that rebalances the weights of the classifier.

5 Conclusion

This work introduces a novel approach to address the problem of multi-modal domain adaptation. Our method is motivated by the observation that differences in the marginal distributions of modalities can significantly affect the training process, leading to suboptimal performance and imbalances in feature norms. To tackle these issues, we introduced the Relative Norm Alignment (RNA) loss, which aims to balance the norms of features extracted by the network across different domains and modalities to improve overall accuracy. This loss is combined with adversarial domain loss and Information Maximization in UDA settings to enhance feature transferability and regularization in the target domain. Our experimental results have shown that the proposed RNA approach can either outperform or compete with several state-of-the-art methods in various multi-modal classification tasks, demonstrating its effectiveness and flexibility. Most notably, our approach is characterized by its simplicity and lightweight nature, allowing it to be easily integrated into different architectures and contexts without requiring complex modifications. This inherent adaptability makes RNA a promising solution for real-world applications where multi-modal data is prevalent. Future research will further explore RNA's capabilities and adaptability to diverse domains and modalities, addressing challenges in unbalanced data distributions, and investigating the integration with other techniques for domain shift mitigation and improved generalization.

Acknowledgements. This work was supported in part by the CINI Consortium through the VIDESEC project, in part by FAIR - Future Artificial Intelligence Research, and in part by European Union Next-Generation EU through PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, under Grant PE00000013. We acknowledge the CINECA award under the ISCRA initiative and IIT HPC infrastructure, for the availability of high performance computing resources and support. The authors would like to thank Prof. Olga Dal Monte for useful discussions on human multimodal perception.

References

[1] P. BERTELSON and B. D. GELDER, “The psychology of multimodal perception,” in Crossmodal Space and Crossmodal Attention. Oxford University Press, Apr. 2004, pp. 141–177.

[2] B. E. Stein, P. J. Laurienti, M. T. Wallace, and T. R. Stanford, “Multisensory integration,” in Encyclopedia of the Human Brain. Elsevier, 2002, pp. 227–241.

[3] J. Driver and C. Spence, “Attention and the crossmodal construction of space,” Trends in Cognitive Sciences, vol. 2, no. 7, pp. 254–262, Jul. 1998.

[4] C. O'Callaghan, Perception and Multimodality. Oxford University Press, May 2012.

[5] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, “Deep audio-visual learning: A survey,” International Journal of Automation and Computing, vol. 18, no. 3, pp. 351–376, 2021.

[6] J. Munro and D. 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.

[7] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, “On the integration of optical flow and action recognition,” in German conference on pattern recognition. Springer, 2019, pp. 281–297.

[8] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. 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, 2018, pp. 1390–1399.

[9] M. R. Loghmani, L. Robbiano, M. Planamente, K. Park, B. Caputo, and M. Vincze, “Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition,” IEEE Robotics and Automation Letters, vol. 5, no. 4, pp. 6631–6638, 2020.

[10] L. Bo, X. Ren, and D. Fox, “Depth kernel descriptors for object recognition,” in 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2011, pp. 821–826.

[11] W. Wang, D. Tran, and M. Feiszli, “What makes training multi-modal classification networks hard?” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12 695–12 705.

[12] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019.

[13] M. Yang, Y. Li, P. Hu, J. Bai, J. Lv, and X. Peng, “Robust multi-view clustering with incomplete information,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 1055–1069, 2022.

[14] P. Razzaghi, K. Abbasi, M. Shirazi, and N. Shabani, "Modality adaptation in multimodal data," Expert Systems with Applications, vol. 179, p. 115126, Oct. 2021.

[15] J. Lv, K. Liu, and S. He, "Differentiated learning for multi-modal domain adaptation," in Proceedings of the 29th ACM International Conference on Multimedia. ACM, Oct. 2021.

[16] M. Planamente, C. Plizzari, E. Alberti, and B. Caputo, "Domain generalization through audio-visual relative norm alignment in first person action recognition," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2022, pp. 1807–1818.

[17] F. Barbato, M. Toldo, U. Michieli, and P. Zanuttigh, "Latent space regularization for unsupervised domain adaptation in semantic segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2021, pp. 2835–2845.

[18] Q. Zhou, W. Zhou, S. Wang, and Y. Xing, "Unsupervised domain adaptation with adversarial distribution adaptation network," Neural Computing and Applications, vol. 33, no. 13, pp. 7709–7721, Nov. 2020.

[19] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5089–5097.

[20] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Los Alamitos, CA, USA: IEEE Computer Society, nov 2019, pp. 1426–1435.

[21] A. Sahoo, R. Shah, R. Panda, K. Saenko, and A. Das, "Contrast and mix: Temporal contrastive video domain adaptation with background mixing," in Thirty-Fifth Conference on Neural Information Processing Systems, 2021.

[22] X. Song, S. Zhao, J. Yang, H. Yue, P. Xu, R. Hu, and H. Chai, "Spatio-temporal contrastive domain adaptation for action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9787–9795.

[23] D. Kim, Y.-H. Tsai, B. Zhuang, X. Yu, S. Sclaroff, K. Saenko, and M. Chandraker, "Learning cross-modal contrastive features for video domain adaptation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13 618–13 627.

[24] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," Information Fusion, vol. 81, pp. 203–239, May 2022.

[25] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5400–5409.

[26] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," JMLR, vol. 13, no. 1, pp. 723–773, 2012.

[27] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in ECCV, 2016.

[28] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," The journal of machine learning research, vol. 17, no. 1, pp. 2096–2030, 2016.

[29] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in Advances in neural information processing systems, 2018, pp. 5334–5344.

[30] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang, "Exact feature distribution matching for arbitrary style transfer and domain generalization," in CVPR, 2022.

[31] C. Chen, J. Li, X. Han, X. Liu, and Y. Yu, "Compound domain generalization via meta-knowledge encoding," in CVPR, 2022.

[32] C. Chen, L. Tang, F. Liu, G. Zhao, Y. Huang, and Y. Yu, "Mix and reason: Reasoning over semantic topology with data mixing for domain generalization," in NeurIPS, 2022.

[33] Y. Wang, H. Li, and A. C. Kot, "Heterogeneous domain generalization via domain mixup," in ICASSP, 2020.

[34] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in AAAI, 2020.

[35] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," in NeurIPS, 2018.

[36] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," Advances in Neural Information Processing Systems, vol. 32, pp. 6450–6461, 2019.

[37] Y. Li, Y. Yang, W. Zhou, and T. Hospedales, "Feature-critic networks for heterogeneous domain generalization," in ICML, 2019.

[38] S. Bucci, A. D'Innocente, Y. Liao, F. M. Carlucci, B. Caputo, and T. Tommasi, "Self-supervised learning across domains," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 9, pp. 5516–5528, 2021.

[39] Z. Yao, Y. Wang, J. Wang, P. Yu, and M. Long, "Videodg: Generalizing temporal relations in videos to novel domains," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2021.

[40] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3723–3732.

[41] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in International conference on machine learning. PMLR, 2015, pp. 97–105.

[42] Z. Deng, Y. Luo, and J. Zhu, "Cluster alignment with a teacher for unsupervised domain adaptation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9944–9953.

[43] H. Tang and K. Jia, "Discriminative adversarial domain adaptation." in AAAI, 2020, pp. 5940–5947.

[44] N. Agarwal, Y.-T. Chen, B. Dariush, and M.-H. Yang, "Unsupervised domain adaptation for spatio-temporal action localization," arXiv preprint arXiv:2010.09211, 2020.

[45] M.-H. Chen, B. Li, Y. Bao, G. AlRegib, and Z. Kira, "Action segmentation with joint self-supervised temporal domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9454–9463.

[46] M.-H. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, and J. Zheng, "Temporal attentive alignment for large-scale video domain adaptation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6321–6330.

[47] J. Choi, G. Sharma, M. Chandraker, and J.-B. Huang, "Unsupervised and semi-supervised domain adaptation for action recognition from drones," in The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 1717–1726.

[48] A. Jamal, V. P. Namboodiri, D. Deodhare, and K. Venkatesh, "Deep domain adaptation in action space," in BMVC, 2018.

[49] B. Pan, Z. Cao, E. Adeli, and J. C. Niebles, "Adversarial cross-domain action recognition with co-attention," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 11 815–11 822.

[50] J. Choi, G. Sharma, S. Schulter, and J.-B. Huang, "Shuffle and attend: Video domain adaptation," in European Conference on Computer Vision. Springer, 2020, pp. 678–695.

[51] F. Qi, X. Yang, and C. Xu, "A unified framework for multimodal domain adaptation," in Proceedings of the 26th ACM international conference on Multimedia. ACM, Oct. 2018.

[52] W. Liu, Z. Luo, Y. Cai, Y. Yu, Y. Ke, J. M. Junior, W. N. Gonçalves, and J. Li, "Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 176, pp. 211–221, Jun. 2021.

[53] M. Jaritz, T.-H. Vu, R. de Charette, E. Wirbel, and P. Pérez, "xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation," in CVPR, 2020.

[54] L. Yang, Y. Huang, Y. Sugano, and Y. Sato, "Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14 722–14 732.

[55] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1041–1049.

[56] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," arXiv preprint arXiv:1703.09507, 2017.

[57] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "COMIC: Multi-view clustering without parameter selection," in Proceedings of the 36th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 09–15 Jun 2019, pp. 5092–5101.

[58] J. Yang, H. Qian, H. Zou, and L. Xie, "Learning decomposed hierarchical feature for better transferability of deep models," Inf. Sci. (Ny), vol. 580, pp. 385–397, Nov. 2021.

[59] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably)," in Proceedings of the 39th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 17–23 Jul 2022, pp. 9226–9259.

[60] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8238–8247.

[61] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1426–1435.

[62] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," ser. Proceedings of Machine Learning Research, vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1180–1189.

[63] X. Wang, L. Li, W. Ye, M. Long, and J. Wang, "Transferable attention for domain adaptation," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 5345–5352, Jul. 2019.

[64] P. Wei, L. Kong, X. Qu, X. Yin, Z. Xu, J. Jiang, and Z. Ma, "Unsupervised video domain adaptation: A disentanglement perspective," arXiv preprint arXiv:2208.07365, 2022.

[65] J. Bridle, A. Heading, and D. MacKay, “Unsupervised classifiers, mutual information and 'phantom targets,” in Advances in Neural Information Processing Systems, vol. 4. Morgan-Kaufmann, 1991.

[66] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price et al., “Scaling egocentric vision: The epic-kitchens dataset,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 720–736.

[67] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, “Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100,” IJCV, 2022.

[68] K. Lai, L. Bo, X. Ren, and D. Fox, “A large-scale hierarchical multi-view rgb-d object dataset,” in 2011 IEEE international conference on robotics and automation. IEEE, 2011, pp. 1817–1824.

[69] M. Papakostas, A. Rajavenkatanarayanan, and F. Makedon, “Cogbeacon: A multi-modal dataset and data-collection platform for modeling cognitive fatigue,” Technologies, vol. 7, no. 2, p. 46, 2019.

[70] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” arXiv preprint arXiv:1212.0402, 2012.

[71] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “Hmdb: a large video database for human motion recognition,” in 2011 International conference on computer vision. IEEE, 2011, pp. 2556–2563.

[72] M. Cannici, C. Plizzari, M. Planamente, M. Ciccone, A. Bottino, B. Caputo, and M. Matteucci, “N-rod: A neuromorphic dataset for synthetic-to-real domain adaptation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1342–1347.

[73] F. Lange, C. Brückner, A. Knebel, C. Seer, and B. Kopp, “Executive dysfunction in parkinson’s disease: A meta-analysis on the wisconsin card sorting test literature,” Neuroscience & Biobehavioral Reviews, vol. 93, pp. 38–56, Oct. 2018.

[74] C. Plizzari, M. Planamente, E. Alberti, and B. Caputo, “Polito-iit submission to the epic-kitchens-100 unsupervised domain adaptation challenge for action recognition,” arXiv preprint arXiv:2107.00337, 2021.

[75] M. Planamente, G. Goletto, G. Trivigno, G. Averta, and B. Caputo, “Polito-iit-cini submission to the epic-kitchens-100 unsupervised domain adaptation challenge for action recognition,” arXiv preprint arXiv:2209.04525, 2022.

[76] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, “Epic-fusion: Audio-visual temporal binding for egocentric action recognition,” in The IEEE International Conference on Computer Vision (ICCV), October 2019.

[77] A. Zisserman, J. Carreira, K. Simonyan, W. Kay, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, and M. Suleyman, “The kinetics human action video dataset,” 2017.

[78] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, “Temporal relational reasoning in videos,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 803–818.

[79] X. Gong, S. Mohan, N. Dhingra, J.-C. Bazin, Y. Li, Z. Wang, and R. Ranjan, “Mmg-ego4d: Multi-modal generalization in egocentric action recognition,” 2023.

[80] D. Damen, E. Kazakos, W. Price, J. Ma, and H. Doughty, “Epic-kitchens-55 - 2020 challenges report,” 2020.

[81] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, “Adaptive batch normalization for practical domain adaptation,” Pattern Recognition, vol. 80, pp. 109–117, 2018.

[82] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in European conference on computer vision. Springer, 2016, pp. 20–36.

[83] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[84] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, pp. 448–456.

[85] 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, 2009, pp. 248–255.

[86] C.-F. R. Chen, R. Panda, K. Ramakrishnan, R. Feris, J. Cohn, A. Oliva, and Q. Fan, "Deep analysis of cnn-based spatio-temporal representations for action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6165–6175.

[87] H. Rebecq, D. Gehrig, and D. Scaramuzza, "Esim: an open event camera simulator," in Conference on robot learning. PMLR, 2018, pp. 969–982.

[88] X. Wu, S. Zhang, Q. Zhou, Z. Yang, C. Zhao, and L. J. Latecki, "Entropy minimization versus diversity maximization for domain adaptation," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–12, 2021.

[89] A. Aakerberg, K. Nasrollahi, and T. Heder, "Improving a deep learning based rgb-d object recognition model by ensemble learning," in 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2017, pp. 1–6.

[90] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[91] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5982–5991.

[92] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," Neural Netw., vol. 106, p. 249–259, oct 2018.

[93] Y. Guo and L. Zhang, "One-shot face recognition by promoting underrepresented classes," 2017.

[94] B. Kim and J. Kim, "Adjusting decision boundary for class imbalanced learning," IEEE Access, vol. 8, pp. 81 674–81 685, 2020.

[95] Y. Wu, H. Liu, J. Li, and Y. Fu, "Deep face recognition with center invariant loss," in Proceedings of the on Thematic Workshops of ACM Multimedia 2017, ser. Thematic Workshops '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 408–414.

[96] M. Li, Y.-M. Cheung, and J. Jiang, "Feature-balanced loss for long-tailed visual recognition," in 2022 IEEE International Conference on Multimedia and Expo (ICME), 2022, pp. 1–6.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available in the following repositories:

EK55: https://data.bris.ac.uk/data/dataset/3h91syskeag572hl6tvuovwv4d

EK100: http://dx.doi.org/10.5523/bris.2g1n6qdydwa9u22shpxqzp0t8m

UCF: https://www.crcv.ucf.edu/data/UCF101.php

HMDB: https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

ROD: http://tiny.cc/NRODDatasetDownload

CogBeacon: https://github.com/MikeMpapa/CogBeacon-MultiModal_Dataset_for_Cognitive_Fatigue