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

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

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

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

Abstract

This supplementary material is divided into eight sections. The first section provides additional details and ablation experiments on the design decisions underlying the Relative Norm Alignment (RNA) loss formulation. In the second section, we explain the effects of minimizing the angular distance between features instead of their norms (as in $\mathcal{L}_{RNA}$). We then show the effects of norm balancing at the class level. Further experiments follow to prove that the benefits of RNA-Net are independent of the fusion strategy and backbone used. We then present additional details on the EPIC -Kitchens-100 setting, squeeze-and-excite adaptation, Non-Local gating [19], and transformer-based approaches [3, 13] referenced in the main paper, including an illustration of the architecture used in the unsupervised (UDA) setting. Finally, we present additional qualitative results.

1. RNA additional details

1.1. RNA^{sub}

An alternative formulation of RNA, in a way similar to HNA, is the "subtraction" as follows:

$$\mathcal{L}_{RNA}^{sub} = (\mathbb{E}[h(X^a)] - \mathbb{E}[h(X^v)])^2. \quad (1)$$

While even this formulation does not introduce any additional hyperparameter, it suffers like HNA from a potentially significant high loss value, which results in highly sensitivity to λ (Figure 7 of main paper).

1.2. Additional Ablation on λ variations

We illustrate in Figure 1 how performances vary based on the weight loss λ assigned to $\mathcal{L}_{RNA}$. We show the variations in terms of accuracy based on the value of λ on all single-source DG, multi-source DG and supervised settings, comparing the results against the standard source-only, DeepAll and supervised baseline respectively. $\mathcal{L}_{RNA}$ outperforms the baseline results on all choices of λ, demonstrating that our loss is not sensitive to λ variations and confirming our choice of the dividend/divisor structure.

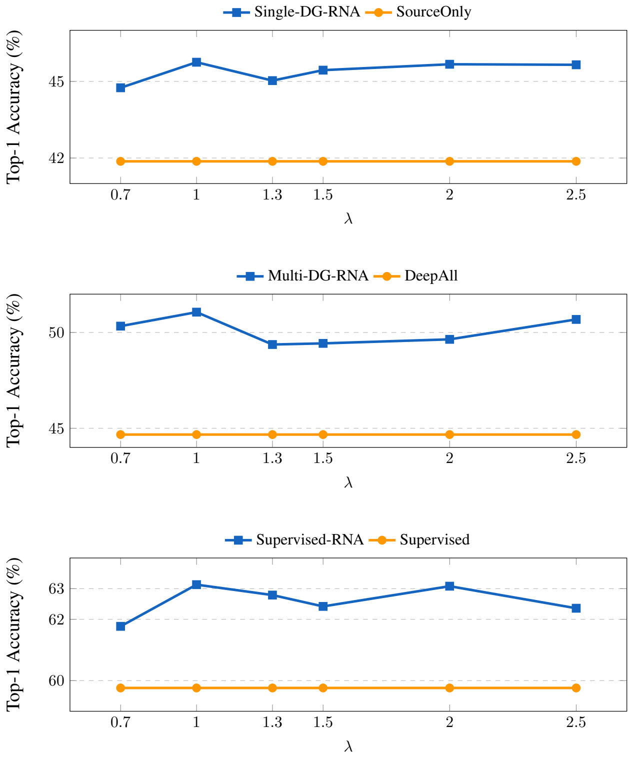

Figure 1. Difference in terms of performance (average Top-1 Accuracy (%)) based on the value of λ used to weight $\mathcal{L}_{RNA}$ on single-DG, multi-DG and supervised setting.

1.3. Temporal Window Ablation

In our experiments, audio length ≈ RGB length, since the 16 frames are strided by 2 as in [14]. We show in the Ta-

[*] The authors equally contributed to this work.

	Audio Length	D1,D2→D3	Audio Norm
DeepAll	[0.64s]	52.19 %	29.92
RNA-Net	[0.64s]	54.31 %	6.80
DeepAll	[2.56s]	51.60 %	30.91
RNA-Net	[2.56s]	53.18 %	6.13
DeepAll	[1.28s]	51.47 %	30.16
RNA-Net	[1.28s]	55.88 %	5.92

Table 1. Comparison in terms of accuracy (%) and mean feature norms between DeepAll and RNA-Net with different audio lengths.

ble 1 how performance and norms vary in correspondence to variations of audio length. The improvement over the baseline is guaranteed in all configurations, confirming that the norm unbalance does not depend on the temporal window.

2. Comparison with other losses

This section details the experiments reported in Table 2 and aims at analyzing the effect of other losses that minimize the angular distance between features rather then their norms.

We refer to these approaches as *Cosine Similarity Loss*, *Orthogonality Loss* and *Euclidean loss* (we refer to them as *CosSim*, *Orth.Only* and *MSE* in Table 2), and we describe them below. All these losses act on the feature vectors f^v and f^a of the two modalities.

The first two are based on a *cosine similarity* metric, i.e., $cos(\theta_{a,v})$ with $\theta_{a,v}$ the angle between f^v and f^a. Their main difference is that CosSim imposes an alignment constraint by minimizing the term $1 - cos(\theta_{a,v})$, thus, forcing the $\theta_{a,v}$ to be zero, while Orth.Only minimizes the item $cos(\theta_{a,v})^2$, forcing the angle $\theta_{a,v}$ to be ±90° (Figure 2) and thus imposing an orthogonality constraint. Instead, the *Euclidean loss* is defined in terms of the *Mean Squared Error (MSE)* between f^v and f^a, thus resulting in the minimization of both $\theta_{a,v}$ and the norm difference between the two feature vectors (Figure 2).

We underline that our $\mathcal{L}_{RNA}$ loss directly operates on feature norms, without imposing any explicit constraint on the angular distance $\theta_{a,v}$ between the two feature vectors. This choice let the network free to choose to exploit modality-specific characteristics, similarly to what happens with Orth.Only, align them, as CosSim and MSE do, or take different orientations.

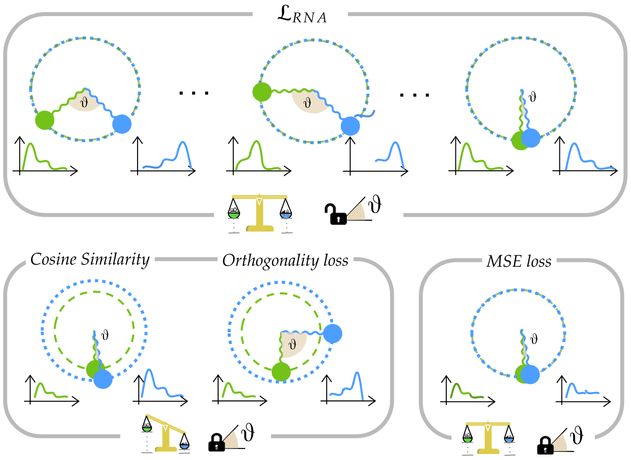

Figure 2. **RNA loss *vs* standard losses.** Representation of visual and audio features by means of segments of different lengths arranged in a radial pattern, and whose length represent their mean feature norm. Similarity-based losses (*Cosine Similarity loss* and *Orthogonality loss*) impose a constraint on the angle θ between the two modality representations, while they do not re-balance feature norms. On the other side, the *MSE loss* minimizes both the angle θ between the two, and the discrepancy between their feature norms. As opposed to this losses, RNA does not impose a constraint on the angular distance θ.

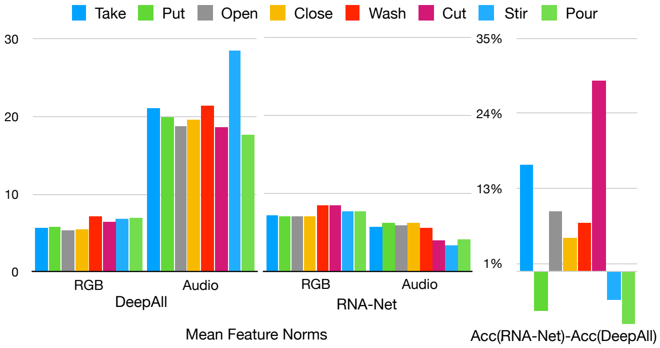

Figure 3. Mean feature norms for each class (left), and the corresponding per-class accuracy (right).

3. Class re-balance

In Figure 3 (left) we report the mean feature norms *for each class*, which has been calculated as a mean of the norm of all samples in the class and mediated over all domains. As it can be observed, the norm unbalance reflects also at class level, with the audio modality's norm still being greater then the visual one *on average*. When minimizing RNA loss, norms are re-balanced not only among modalities, but also across all classes. In Figure 3 (right), we also show the improvement in terms of per-class accuracy of RNA-Net w.r.t. the baseline DeepAll. In general, accuracy improves across basically all classes. This is because the RNA loss prevents the audio modality (higher norm) from "dominating" the visual modality (lower norm), allowing the network to accurately leverage knowledge from the visual stream in classifying actions.

	Comparison with other losses										
	D1 → D2	D1 → D3	D2 → D1	D2 → D3	D3 → D1	D3 → D2	Mean	D1, D2 → D3	D1, D3 → D2	D2, D3 → D1	Mean
CosSim	38.50	33.75	32.59	45.78	39.97	50.86	41.76	50.01	42.40	44.40	45.60
Orth. Only	39.18	37.55	36.86	47.09	43.70	51.61	42.67	53.08	41.76	48.07	47.64
MSE	43.88	44.18	40.28	49.89	43.27	51.60	45.52	40.80	46.59	50.93	46.11
RNA loss (Ours)	45.01	44.62	41.76	48.90	42.20	51.98	**45.75**	55.88	45.65	51.64	**51.06**

Table 2. Top-1 Accuracy (%) of RNA loss w.r.t. to other losses, namely the cosine similarity, an orthogonality loss, and MSE, on both single- and multi-DG settings.

	Fusion strategies			
	Supervised	Single-DG	Multi-DG	DA
Baseline (*mid-level fusion*)	60.18	40.33	47.61	40.33
RNA-Net (*mid-level fusion*)	62.11	45.48	49.56	45.73
Baseline (*late fusion*)	59.76	40.93	44.67	40.93
RNA-Net (*late fusion*)	**63.13**	**45.75**	**51.06**	**47.71**

Table 3. Top-1 Accuracy (%) of RNA-Net with two different fusion strategies, namely *mid-level fusion* and *late fusion*. **Bold:** highest result for each setting.

4. Architectural Variations

Fusion Strategies. In Table 3 we compare the *late fusion* approach against the so-called *mid-level fusion*, used in [10]. It consists in feeding the prediction layer with the fusion of the two modality features, i.e., concatenation. As far as it concerns the mid-level fusion approach, it demonstrates to be a valid alternative in both the intra-domain (*Supervised*) and cross-domain scenarios, remarking the flexibility of our method to be employed with different multi-modal fusion strategies and in all the presented settings.

TBN Variation. To prove that the RNA loss could be easily included in other networks, we report some results obtained by integrating RNA loss in the original TBN [10] network. Those are: D1,D2→D3: 52.69 (+3.8%), D1,D3→ D2: 46.16 (+3.5%) , D2,D3 → D1: 46.68 (-0.7%), avg: 48.51 (+2.2%). The improvement is reported w.r.t. to Deep All results in the main paper.

5. EPIC-Kitchens-100 UDA Setting

The EPIC-Kitchens-100 dataset consists of first-person videos of 16 participants in several kitchens and 97 verb classes. The UDA setting provides as source the videos recorded in 2018 (EK-55) and as target data (unlabeled) videos recorded 2 years later (2020). The temporal gap, also referred to as temporal domain shift, is the difficulty that the domain adaptation techniques have to face.

The multi-source nature of the proposed setting makes it perfect to deal with the domain shift using DG techniques. Moreover, we believe that a domain generalization algorithm serves precisely to ensure the robustness of a model over time, without having to resort to UDA-like approaches that require the use of the (albeit unlabelled) target.

Authors of [4] proposed a multi-modal (RGB+A+OF) version of **Temporal Attentive Adversarial Adaptation Network (TA^{3}N) [2]** as UDA baseline. The UDA technique proposed in [2] is based on three components. The first one, called *Temporal Adversarial Adaptation Network (TA^{2}N)*, consists in an extension of DANN [6], aiming to align the temporal features on a multi-scale Temporal Relation Module (TRM) [20] through a gradient reversal layer (GRL). The second component is based on a domain attention mechanism which guides the temporal alignment towards features where the domain discrepancy is larger. Finally, the third component uses a minimum entropy regularization (attentive entropy) to refine the classifier adaptation.

To ensure a fair comparison, we added optical flow into our architecture (using the same features extractor that was used in [4]) while still minimizing the RNA loss. As it can be observed from Table 6 of the main paper, even without employing the target, we outperform TA^{3}N, confirming the importance of DG in this context. We also demonstrated that RNA-Net and TA^{3}N approaches are complementary.

6. Implementation Details

This section details the adaptions applied to the implementation of Squeeze-and-Excitation (Section 6.1) and Non-Local (Section 6.2) to use them as gating mechanism (following [18]), and the Transformer-based approaches (Section 6.3). Finally, we provide additional details on how we adapted other multi-modal and UDA approaches to our framework.

6.1. Squeeze And Excitation (SE) Gate

The *Squeeze-and-Excitation* networks [7] introduce a gating mechanism aimed at selectively adjusting each feature map's weighting according to the task at hand. We adapted this approach to our multi-modal network using the input features from one stream to compute, using the same recipe of [7], scaling weights for the features of the other stream. Specifically, after performing a global average pooling on the features from the visual stream, we learn weights as in [7], and use them to scale the features of the audio stream. The same procedure is applied on audio features to produce weights for the visual ones. The so obtained scaled channels are then used to replace the existing ones of the corresponding modality.

We found this mechanism to perform the best when added after the third block of both the BNInception [8] and Inflated Conv3D (I3D) [1] models.

6.2. Non-Local (NL) Gate

Non-Local Networks [19] proposed an attention-based gating mechanism, which follows a Query-Key-Value formulation. When gating from the visual stream to the audio stream, the Query is calculated from the visual stream features and Key and Value from the audio one (and vice versa). Both implementations are shown in figure 5. Similarly to the SE gate, we applied non-local gating after the audio and visual models' third block.

6.3. Transformer-based Approaches

In our work, we adopted the two following recent transformer-based approaches proposed as a co-attention mechanism to learn generic audio-visual representations:

- the Transformer Self-Attention [13], which leverages a translation network that first combines a set of features from k audio or visual clips (extracted before the final global pooling) using Query-Key-Value Self-Attention [17]; then, a second module computes a simple clip-wise feed-forward transformation (figure 6-a) applied, in our implementation, to both audio (f^a) and visual (f^v) features.

- the Cross-Modal Transformer Self-Attention [3], which modifies the queries of each modality using Self-Attention on the features of the other modality (figure 6-b).

6.4. Other implementation details

We provide additional details on the implementation of the multi-modal and UDA approaches illustrated in Table 1-2 of the main paper. Notice that, for a fair comparison, we kept unaltered our backbone architectures in all of these experiments, except for TBN [10], for which we maintained the original backbone architecture.

- IBN-Net [15]: we combined *Instance Normalization* [16] and *Batch Normalization* [9] layers as in [15];

- Gradient Blending [18]: we re-weighted the visual and RGB losses, and their joint combination loss as in [18], using the weights used in the original paper [18] and provided in the EPIC-Kitchens Action Recognition challenge report for the verb category [5];

- UDA approaches: we adapted AdaBN [11] by updating separate batch normalize layers with source and target statistics respectively. As far as GRL [6], we added a gradient reversal layer on top of the features extractor and trained a domain discriminator. Finally, we integrated MMD loss in [12] on our framework.

7. RNA-Net architecture for UDA

Thanks to RNA loss's unsupervised nature, we can extend RNA-Net to the Unsupervised Domain Adaptation (UDA) setting, where unlabelled target data are available during training. As shown in figure 4 this can be done by simply redefining $\mathcal{L}_{RNA}$ as the sum of $\mathcal{L}^s_{RNA}$ and $\mathcal{L}^t_{RNA}$, where each term computes a loss on different domains (i.e., respectively, source and target). The action classifiers G^v and G^a are then trained with standard cross-entropy loss on source data only and RNA-Net is trained by minimizing the following total loss:

$$\mathcal{L} = \mathcal{L}_C + \lambda(\mathcal{L}^s_{RNA} + \mathcal{L}^t_{RNA}), \tag{2}$$

where λ indicates the weight given to our loss $\mathcal{L}_{RNA}$.

8. Additional Qualitative Results

To conclude these supplementary, we present some additional qualitative results based on the Class Activation Maps obtained by training the network with and without the additional RNA loss. The images in Figure 7 clearly show that, by rebalancing the modalities norms, the network improves its ability to correctly identify the image regions that correlate best with the represented action. This effect is due to the fact that when it unifies the norms, the network also tends to "choose" which features to favour (i.e., those that are more general) and, consequently, to reduce those that it considers irrelevant, i.e., those more domain-specific (such as the background) that negatively influence generalization.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[2] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6321–6330, 2019.

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

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020.

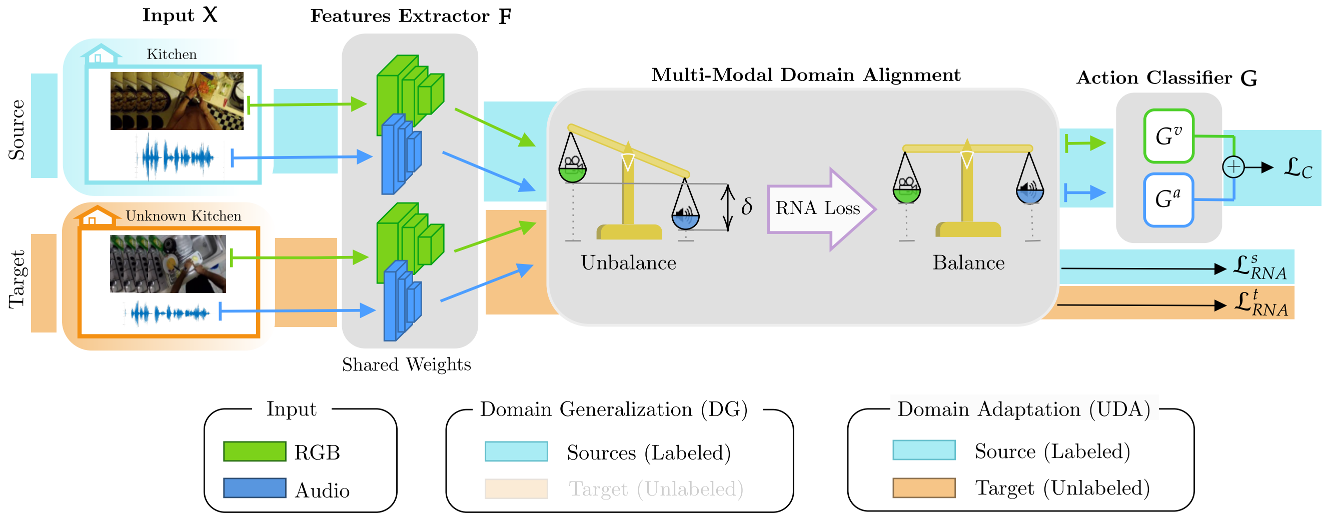

Figure 4. **RNA-Net.** Labeled source visual $x^v_{s,i}$ and source audio $x^a_{s,i}$ inputs are fed to the respective feature extractors F^v and F^a. Unlabeled target data of any modality $(x^m_{t,i})$ is seen at training time only in UDA setting, and not in DG. Our $\mathcal{L}_{RNA}$ operates at feature-level by balancing the relative feature norms of the two modalities, on both source and target data ($\mathcal{L}^s_{RNA}$ and $\mathcal{L}^t_{RNA}$ respectively). The action classifiers G^v and G^a are trained with standard cross-entropy loss $\mathcal{L}_c$ on source labeled data only.

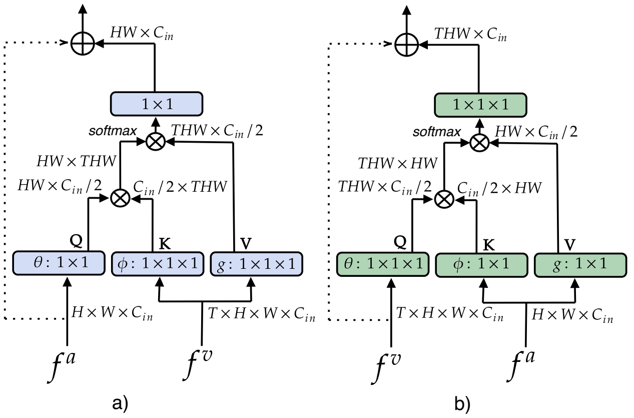

Figure 5. **Non-local Gating.** Left: NL-Gate on audio steam. Right: NL-Gate on visual stream.

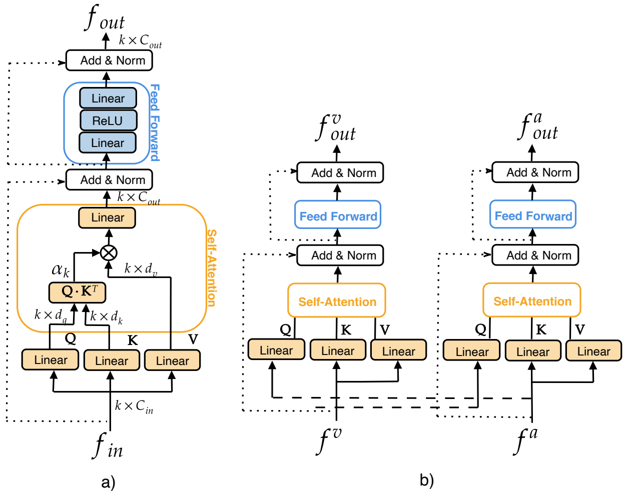

Figure 6. **Transformer-based Approaches.** Left: Transformer architecture for feature translation from [13]. The *query* **Q**, *key* **K** and *value* **V** vectors are learned through linear transformations, and the Query-Key dot product is performed. The resulting *attention weights* α_k are then used to weight the corresponding *value* vectors. The *Feed Forward module* that follows outputs the final features f_{out}. Right: Cross-Modal Attention from [3].

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

[6] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR.

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

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

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

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

[11] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical do-

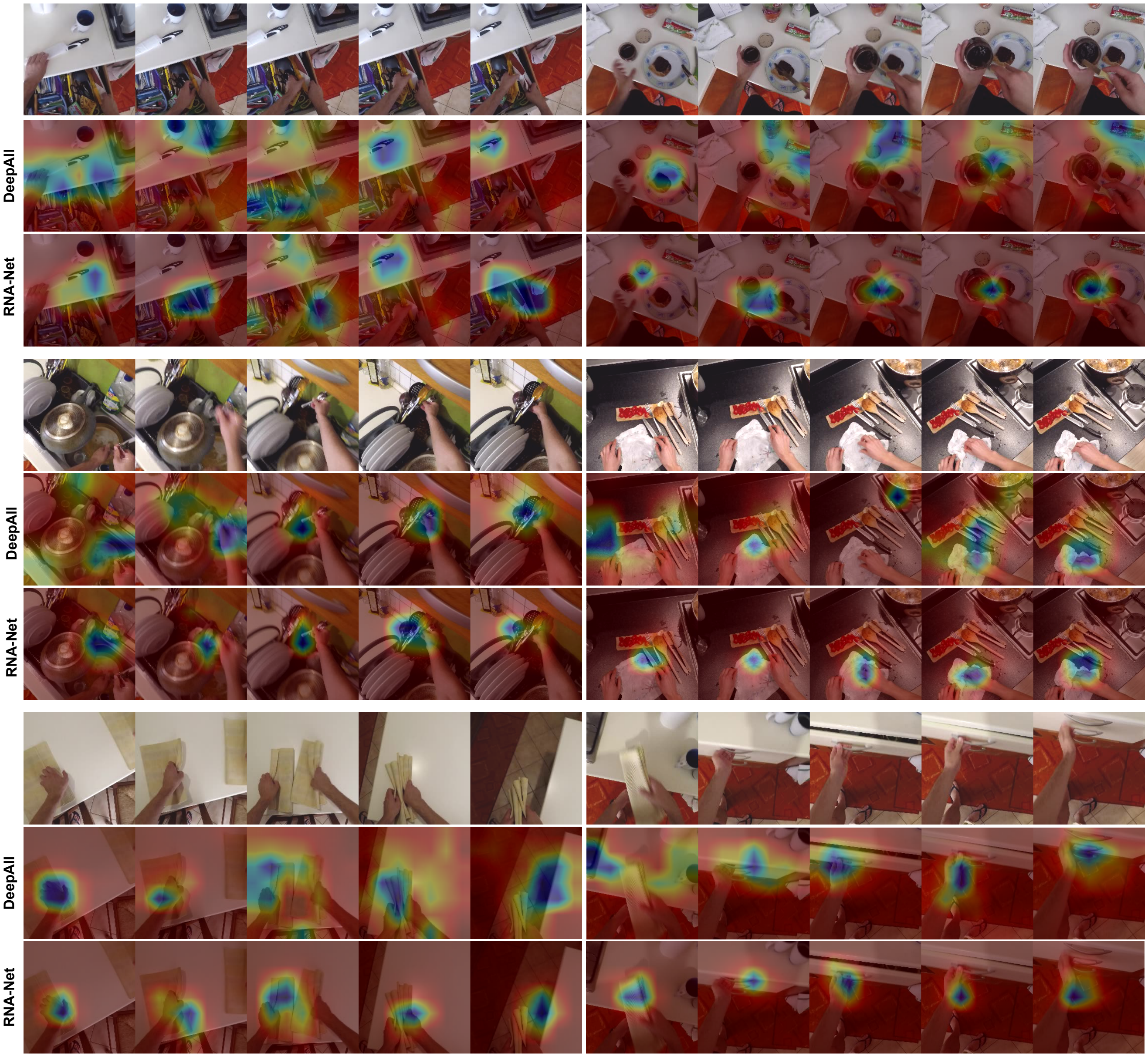

Figure 7. **Class Activation Maps.** Illustration of the CAMs with and without RNA loss. When using RNA loss, the network focuses on the most discriminative part of the image in order to classify the action.

main adaptation. *Pattern Recognition*, 80:109–117, 2018.

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

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

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

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

[16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. pages 4105–4113, 07 2017.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red

Hook, NY, USA, 2017. Curran Associates Inc.

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

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

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