Interact before Align: Leveraging Cross-Modal Knowledge for Domain Adaptive Action Recognition Supplementary Material

Lijin Yang, Yifei Huang, Yusuke Sugano, Yoichi Sato
Institute of Industrial Science, The University of Tokyo
{yang-lj,hyf,sugano,ysato}@iis.u-tokyo.ac.jp

1. Additional implementation details

For all experiments, the MC module processes the feature with $\mathbb{R}^{1024\times7\times7}$ for RGB and Flow modalities, and $\mathbb{R}^{1024}$ for the Audio modality. k is set to 3. The ratio for gating bottleneck is $r = 16$. Dataset-specific details are as follows:

- **U-H dataset**

  We first extract features using I3D [1] pretrained on Kinetics. For each action clip, we extract features from 25 uniformly sampled frames. We use the same strategy as TSN [13] to choose 5 frames from 25 frames. For training our CIA model, we apply Adam optimizer [8] with learning rate 3e-3. We empirically choose $\lambda_y = 1$, $\lambda_{vd} = 1$ and $\lambda_{fd} = 0.5$ for the experiments.

- **E55 dataset**

  On E55 dataset, we train I3D backbone together with our CIA model using Adam optimizer [8] with learning rate 1e-4. We uniformly sample 16 frames as the inputs. We empirically choose $\lambda_y = 1$, $\lambda_{vd} = 1$ and $\lambda_{fd} = 0.5$ for the experiments.

- **E100 dataset**

  For the experiments using I3D as backbone, we apply the same training method as for the E55 dataset.

  For the experiments that use TBN [6] as backbone, we first extract features using TBN fine-tuned on the source dataset following [3]. For each action clip, we extract features from 25 uniformly sampled frames. We use the same strategy as TSN [13] to choose 5 frames from 25 frames. For training the model, we apply Adam optimizer [8] with learning rate 1e-4. Specifically, when using TRN [19] as the temporal aggregation method, we train the model using SGD optimizer with learning rate 3e-3.

2. Analysis on parameters and computational complexity

We show the parameter with and without our proposed CIA model on the I3D backbone in Table 1. The case of two-stream input (RGB and Flow) is shown. Our proposed CIA model introduces a very small amount of additional parameters and computational complexity.

	Number of parameters	Computational complexity
I3D	25.57 M	53.46 GMac
I3D + Ours	27.64 M	53.51 GMac

Table 1. Model parameter and computational complexity.

3. T-SNE visualization

Figure 1 shows the t-SNE [12] visualization of the feature spaces produced by TA^{3}N (a) and TA^{3}N + CIA(ours) (b) on **U-H** dataset. Our CIA increased accuracy of TA^{3}N from 89.17 to 91.94, and the domain alignment is more visible especially in the zoomed in area, showing that our CIA increases feature transferability.

4. Visualization of consensus map

In addition to Grad-CAM [10] visualizations of SC, we directly show the spatial consensus map obtained from SC for more comprehensive understanding on how it works. An example visualization can be found in Fig. 2.

5. Results on E100 test set

Our method ranks on top of the EPIC-KITCHENS-100 2021 challenge leader-board of unsupervised domain adaptation of action recognition. Please refer to our technical report [17] and challenge results [4] for more details.

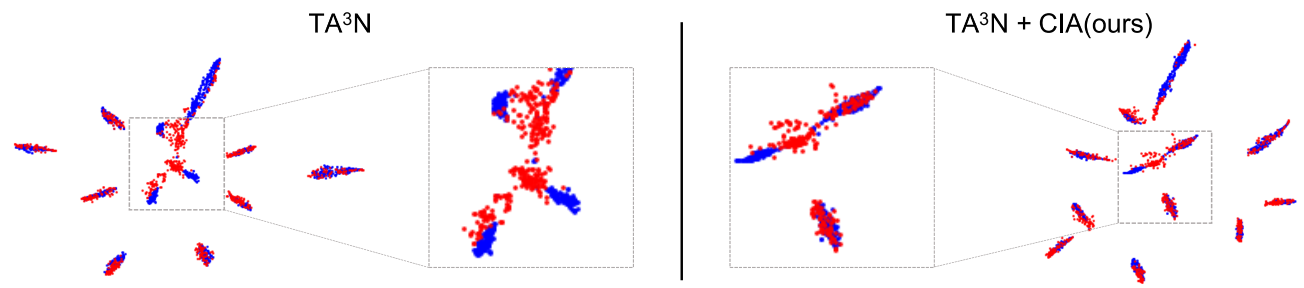

Figure 1. t-SNE plots of feature spaces produced by TA^{3}N (a) and TA^{3}N+CIA (b). Source is shown in blue and target in red. Our method better aligns source and target domains.

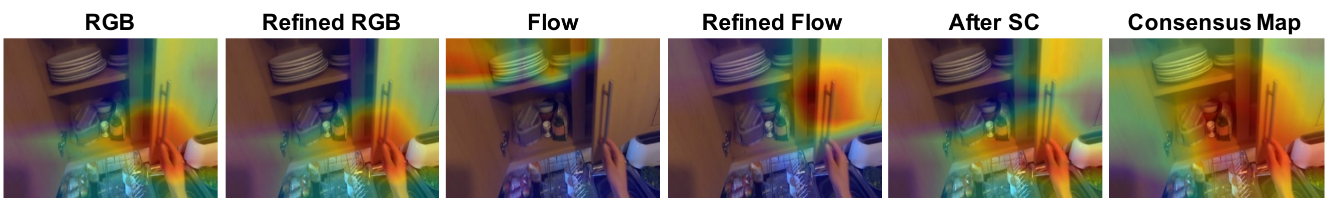

Figure 2. Grad-CAM [10] visualizations of RGB, refined RGB, Flow, refined Flow and fused modality after SC as well as the consensus map obtained by SC. The ground-truth action label is *open cupboard*.

6. More results on different design options of SC

The SC module aims to spatially re-weight the features based on the transferability of each location. We first compare with the most widely used fusion methods: spatial max pooling (**Max**) and average pooling (**Avg**) as well as spatial attention mechanisms for general purpose (**Att**) and for domain adaptation (**TADA**).

Other than using both modalities to generate the spatial map, recent researches [2, 18] found that the Flow modality is stronger in encoding motion information and thus used Flow as the pivot to guide other modalities. We also experiment using a similar setting where we use Flow attention to guide the RGB attention (**Att***) as an additional comparison baseline. In Table 2, Att* is better than Att but worse than our SC. This is because compared with just using the Flow modality to lead the RGB modality, our SC can also use RGB to correct the Flow modality.

We also show the results of our SC with different k value. By comparing the accuracy of verb, noun and action, we can conclude the usefulness of multi-scale correlation.

Module	Verb	Noun	Action
Avg	47.96	29.08	19.19
Max	48.11	29.59	19.48
Att [16]	48.08	29.46	19.39
TADA [15]	47.79	29.69	19.59
Att*	48.29	29.56	19.62
SC ($k = 1$)	48.39	29.70	19.62
SC ($k = 2$)	48.57	29.72	19.77
SC ($k = 3$)	**48.66**	**29.79**	**19.83**

Table 2. Performance comparison of our SC module with other approaches on the **E100** validation set.

7. More results on UCF-HMDB dataset

We show more results on the UCF-HMDB dataset under the source only setting in Table 3. Better representation leads to high source only performance, while larger improvement by DA shows that our features are more transferable. For example, our CIA result is consistently better than MMTM [5] under the same setting. To be noted, our CIA before and after DA enjoys larger gain than MMTM, ours 86.11 to 88.33 (+2.22) vs. MMTM 85.00 to 85.83 (+0.83), which also validates our claim that CIA leads to better transferability.

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. 1

[2] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Con-*

Setting	Method	U→H	H→U
Source only	G-blend [14]	83.33	87.39
	MMTM [5]	85.00	87.74
	STCDA [11]	82.80	89.80
	Kim *et al.* [7]	82.80	90.70
	CIA (Ours)$^{\diamond}$	86.11	92.47
	Concat*	83.89	90.02
	CIA (Ours)*	85.83	93.52
Domain Adaptation	G-blend [14]	84.72	91.24
	MMTM [5]	85.83	92.47
	MM-SADA [9]	84.20	91.10
	STCDA [11]	83.10	92.10
	Kim *et al.* [7]	84.70	92.80
	CIA (Ours)$^{\diamond}$	**88.33**	**94.05**
	Concat*	86.11	92.99
	CIA (Ours)*	**90.56**	**94.22**

Table 3. Performance comparison on the UCF-HMDB (**U-H**)) dataset. $^{\diamond}$ refers to averaging the outputs from each modality classifier, while * means concatenate features of different modalities.

ference on Computer Vision and Pattern Recognition, pages 7882–7891, 2019. 2

[3] 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. 1

[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: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. 1

[5] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13289–13299, 2020. 2, 3

[6] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 1

[7] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021. 3

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[9] 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*, pages 122–132, 2020. 3

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2

[11] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9787–9795, 2021. 3

[12] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[13] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1

[14] 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. 3

[15] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5345–5352, 2019. 2

[16] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[17] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Epic-kitchens-100 unsupervised domain adaptation challenge for action recognition 2021: Team m3em technical report. *arXiv preprint arXiv: 2106.10026*, 2021. 1

[18] Jiaojiao Zhao and Cees GM Snoek. Dance with flow: Two-in-one stream action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9935–9944, 2019. 2

[19] 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. 1