Learning Cross-modal Contrastive Features for Video Domain Adaptation Supplementary Material

Donghyun Kim[1], Yi-Hsuan Tsai[2], Bingbing Zhuang[2], Xiang Yu[2]
Stan Sclaroff[1], Kate Saenko[1,3], Manmohan Chandraker[2]
[1]Boston University, [2]NEC Labs America, [3]MIT-IBM Watson AI Lab
[1]{donhk, sclaroff, saenko}@bu.edu, [2]{ytsai, bzhuang, xiangyu, manu}@nec-labs.com

A. Implementation Details

We use 2 TITANXP GPUs in our implementation. We also reproduce the results of MM-SADA [1][1] using their released code with the same 2-GPU setup and the same batch size as our method.

B. Ablation study

In Table A, we show the sensitivity analysis on the λ (Eq. (7) in the main paper) and the confidence threshold T for pseudo-labels (Section 3.3 in the main paper).

In Table B, we first show the benefit of having the projection head $h(\cdot)$ for multi-modal embedding space. We observe a 2% performance gain by adding the projection head $h(\cdot)$, which demonstrates the importance of using $h(\cdot)$ for multi-modal regularization described in Section 3.2 of the main paper. Moreover, we provide another ablation study where we add the projection head $h(\cdot)$ in the cross-domain module, while having $h(\cdot)$ for the cross-modal module as in our final model. Adding $h(\cdot)$ shows slightly worse results than our final model. One reason is that this scheme has less influence on the features that are supposed to be aligned for performing action recognition.

In Table C, we provide experimental results when different feature alignment methods are used in either cross-modal or cross-domain learning. In general, using the proposed contrastive learning method in both modules obtains the best performance, which shows the importance of having a unified contrastive learning framework for cross-modal and cross-domain learning.

B.1. t-SNE Feature Visualizations

Figure A shows different combinations of the feature spaces before the projection head $h(\cdot)$, *i.e.*, F_s^a, F_s^m, F_t^a, F_t^m. Figure A-(a,b) shows the RGB and flow features in each domain. While the RGB and flow features are almost completely aligned after the projection head in Figure 3 of the main paper, here the RGB and flow features still keep their respective information before the projection head $h(\cdot)$, which is useful for final action predictions. Figures A-(c,d)

[1]https://github.com/jonmun/MM-SADA-code

Table A: Ablation study on hyper-parameters on Epic-Kitchens. In the second group, we fix $T = 0.8$, while in the third group, we fix $\lambda = 1.25$.

Setting	Mean
Source-only	45.5
Ours ($\lambda = 1.25, T = 0.8$)	51.0
Ours ($\lambda = 1.0$)	50.1
Ours ($\lambda = 1.5$)	49.5
Ours ($T = 0.9$)	49.6
Ours ($T = 0.6$)	49.8

Table B: Ablation study on the projection head $h(\cdot)$ for EPIC-Kitchens.

Setting	$h(\cdot)$ in cross-modal module	Mean
Ours (modality)	✓	48.7
Ours (modality)	✗	46.7
Setting	**$h(\cdot)$ in cross-domain module**	**Mean**
Ours (modality + domain)	✓	50.1
Ours-final (modality + domain)	✗	51.0

Table C: Ablation study of different feature alignment methods on EPIC-Kitchens. "Con." indicates our proposed contrastive learning approach, and "Adv." denotes the adversarial learning scheme.

Setting	Modality	Domain	Mean
Con. (our final model)	Con.	Con.	51.0
Adv.	Adv.	Adv.	49.5
Adv. + Con.	Con.	Adv.	50.1

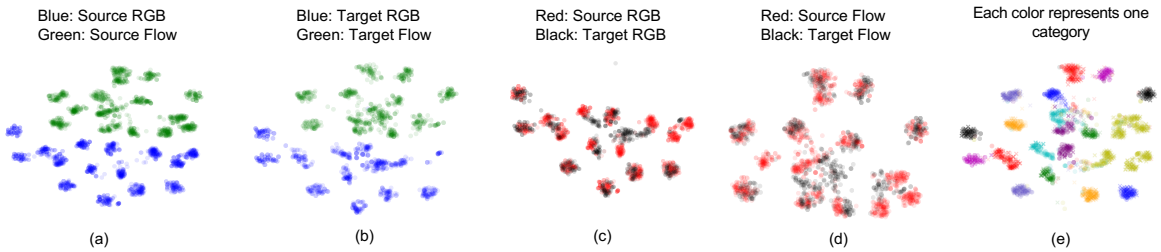


Figure A: t-SNE visualization on cross-modal and cross-domain features before the projection head $h(\cdot)$ on UCF $\rightarrow$ HMDB, *i.e.* F_s^a, F_s^m, F_t^a, F_t^m. In (a)(b), we show the visualization for individual domains, where each domain contains the multi-modal features. In (c)(d), we visualize features for each modality, and each plot uses the features from two domains. (e) includes all the features from two domains and two modalities, where each color represents one action class.

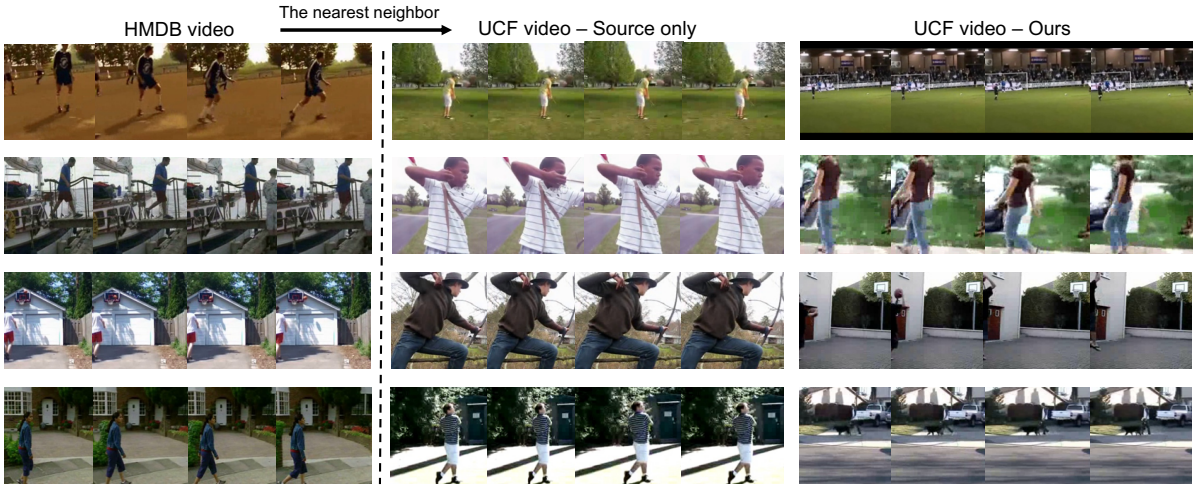
shows how the source and target features are aligned in each modality, where our method can learn domain-invariant features.

B.2. Visualizations for Cross-domain Retrievals

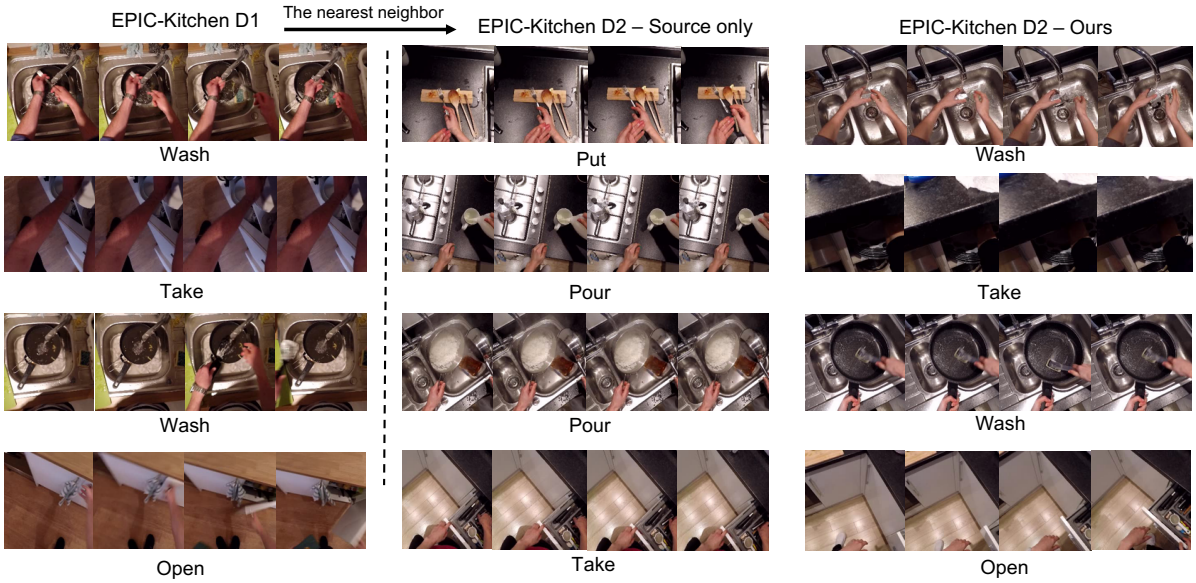
In Figure B-(a), based on the target feature in HMDB, we show the nearest neighbor one from UCF. Similarly, we show the retrievals from EPIC Kitchens D1 and D2 in Figure B-(b). EPIC-Kitchen is more challenging than UCF-HMDB as it has more common background (*e.g.*, similar kitchen backgrounds) or objects (*e.g.*, frying pan, utensils) between different action classes. Our method shows better results that retrieve the videos of the same class.

References

[1] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, 2020. 1



(a) UCF→HMDB



(b) EPIC-Kitchen D1→D2

Figure B: Cross-domain retrievals in the RGB embedding space. Given the target feature F_t^a, we retrieve the closest neighbor F_s^a in the source domain. By our contrastive learning framework, our model correctly aligns videos of the same class, while the source-only model are more likely to be biased to the background context.