# Multimodal Cross-Domain Few-Shot Learning for Egocentric Action Recognition
– Supplementary Materials –

Masashi Hatano[1], Ryo Hachiuma[2], Ryo Fuji[1], and Hideo Saito[1]

[1] Keio University
[2] NVIDIA

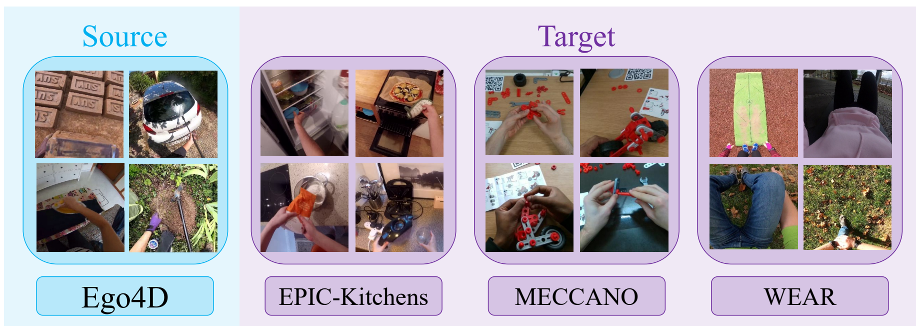
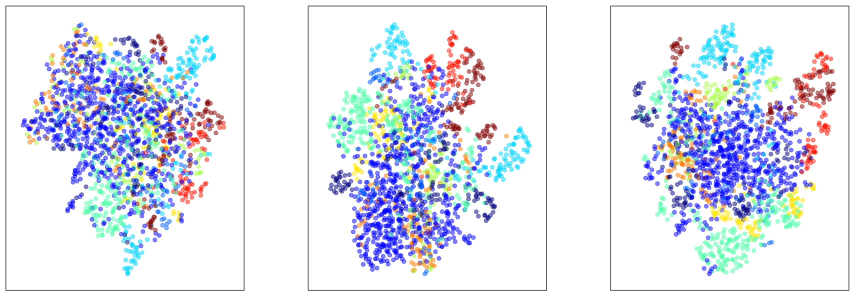



**Fig. 4: Samples from each dataset**. A curated selection of RGB images from each dataset showcases the domain gap between the source and target datasets.

## A Datasets

In this section, we delve into the datasets used in our study, highlighting the significant domain gap observed in RGB images between the source (Ego4D [12]) and target (EPIC-Kitchens [5], MECCANO [43], and WEAR [1]) datasets. Fig. 4 presents a curated selection of RGB images drawn from each dataset. These visual examples underscore the diversity and complexity of cross-domain few-shot learning (CD-FSL) tasks.

– **Ego4D**. The Ego4D dataset comprises a diverse range of activities, including domestic chores in houses, gardening, cleaning, building, cooking, and arts/crafting. These actions contain those performed using a single hand, both hands, and various tools, demonstrating the diversity of human activities. Thus, this dataset is well-suited for learning generalizable features required for cross-domain and few-shot settings.
– **EPIC-Kitchens**. The EPIC-Kitchens dataset, serving as the target, is centered around kitchen activities, encompassing actions like "pouring flour", "opening the refrigerator", "moving a pizza", and "pouring a smoothie", each composed of a verb-noun pair. Such activities in the kitchens are also present within the Ego4D dataset, our source. However, the EPIC-Kitchens dataset distinguishes itself by its fine-grained action categories, with some actions



(a) Only RGB (only reconstruction) (b) Only RGB (c) Ours

**Fig. 5: Comparative UMAP visualization of feature representations**. UMAP plot of 10 classes from EPIC-Kitchens validation set with features obtained from (a) Only RGB (only reconstruction), (b) Only RGB, and (c) Ours.

sharing verbs but differing in nouns and others vice versa. This granular differentiation of actions introduces a challenge for few-shot learning models, necessitating nuanced discernment in action categorization.

- **MECCANO**. The MECCANO dataset comprises detailed recordings of the assembly process for a toy bicycle in an industrial-like scenario. It contains fine-grained actions, including "aligning a screwdriver to screw", "aligning objects", "putting a tire", and "putting a screw". The diminutive size of the components poses a substantial challenge for action recognition, requiring exceptional precision to identify and understand the intricate interactions.
- **WEAR**. This dataset captures outdoor workout actions such as "stretching hamstrings", "jogging", "pushing-ups", and "sitting-ups". Unlike the other datasets, WEAR focuses on activities not involving hand-object interactions but body movements.

## B Visualization of Feature Representations

We compare the class-discriminativeness of embeddings extracted from three encoders: Only RGB (only reconstruction), Only RGB, and Ours. The only RGB (only reconstruction) model is trained without multimodal distillation, and $\lambda_{\mathrm{ce}_{\mathrm{RGB}}} = 0$ is used during the pretraining stage. The only RGB model is trained without multimodal distillation, and $\lambda_{\mathrm{ce}_{\mathrm{RGB}}} = 0.05$ is used during the pretraining stage. Fig. 5 shows the UMAP plot of 10 classes from EPIC-Kitchens datasets. We see that only RGB model creates better grouping on the embeddings of the target datasets than only RGB (only reconstruction) model. This result supports that using the cross-entropy loss helps learn the class-discriminative features during the pretraining stage. We further see that the multimodal distillation also helps learn discriminative features compared to the only RGB model.

## C Loss Weight Ablation

We present an ablation study focused on the impact of adjusting the loss weight for the cross-entropy loss on the RGB modality $\lambda_{\mathrm{ce}_{\mathrm{RGB}}}$ during the pretraining stage. The loss weight for the cross-entropy loss $\lambda_{\mathrm{ce}_{\mathrm{RGB}}}$ serves as a critical hyperparameter that balances the contribution of the cross-entropy loss to the total loss function. For the ablation study, we varied the value of the loss weight across a predefined range $\lambda_{\mathrm{ce}_{\mathrm{RGB}}} \in \{1, 0.1, 0.05, 0.01\}$. We report the 5-way 5-shot action recognition accuracy on the EPIC-Kitchens dataset of the only RGB model with varied hyperparameter $\lambda_{\mathrm{ce}_{\mathrm{RGB}}}$ in Tab. 5. Our ablation analysis reveals a critical insight: a high or low loss weight for the RGB modality, $\lambda_{\mathrm{ceRGB}}$, during the pretraining stage can detrimentally affect the acquisition of class-discriminativeness on the target data. Assigning a high loss weight to the cross-entropy reduces the relative contribution of learning from unlabeled target data. Conversely, a low loss weight fails to capture class discriminativeness adequately.

**Table 5: Loss wight ablation**. We conduct an ablation study on the loss weight for cross-entropy loss on RGB modality pertaining.

| $\lambda_{\mathrm{ce}_{\mathrm{RGB}}}$ | 5-shot |
|---|---|
| 1 | 54.58 |
| 0.1 | 56.68 |
| 0.05 | **57.07** |
| 0.01 | 52.40 |