Multimodal Cross-Domain Few-Shot Learning for Egocentric Action Recognition

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

[1] Keio University
[2] NVIDIA

Abstract. We address a novel cross-domain few-shot learning task (CD-FSL) with multimodal input and unlabeled target data for egocentric action recognition. This paper simultaneously tackles two critical challenges associated with egocentric action recognition in CD-FSL settings: (1) the extreme domain gap in egocentric videos (*e.g.*, daily life vs. industrial domain) and (2) the computational cost for real-world applications. We propose MM-CDFSL, a domain-adaptive and computationally efficient approach designed to enhance adaptability to the target domain and improve inference speed. To address the first challenge, we propose the incorporation of multimodal distillation into the student RGB model using teacher models. Each teacher model is trained independently on source and target data for its respective modality. Leveraging only unlabeled target data during multimodal distillation enhances the student model's adaptability to the target domain. We further introduce ensemble masked inference, a technique that reduces the number of input tokens through masking. In this approach, ensemble prediction mitigates the performance degradation caused by masking, effectively addressing the second issue. Our approach outperformed the state-of-the-art CD-FSL approaches with a substantial margin on multiple egocentric datasets, improving by an average of 6.12/6.10 points for 1-shot/5-shot settings while achieving 2.2 times faster inference speed. Project page: https://masashi-hatano.github.io/MM-CDFSL/

Keywords: Egocentric Vision · Action Recognition · Cross-Domain Few-Shot Learning · Multimodal Distillation

1 Introduction

The field of egocentric vision research, primarily facilitated by wearable devices such as smart glasses [6], has seen significant development over the last decades, owing to its wide range of application domains, including daily life, industry, AR/VR, and medicine. In response to the growing demand across multiple domains, several large-scale egocentric datasets, such as Ego4D [12] and Ego-Exo4D [13], have been proposed to provide a variety of research tasks with various modalities [10,17,40,44,45,56]. Recognizing the action of the camera wearer (*egocentric action recognition*) is one of the fundamental tasks in the egocentric video understanding domain.
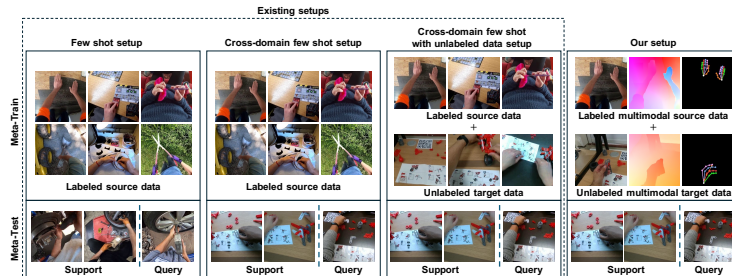
Fig. 1: **Problem setup.** In our problem setup, a model is trained using source data and unlabeled target data for multiple modalities during the meta-training stage. In the meta-testing stage, a few examples of novel classes from the support set are provided to learn a classifier. Then, the network predicts the categories of different samples from the query set, which are the same classes as the support set. Unlike existing setups, we leverage multimodal data (*e.g.*, optical flows or hand poses) during the meta-training stage. During the meta-testing stage, only RGB videos are used as inputs.

Despite these advancements, a scarcity of egocentric datasets persists in various domains, such as industry and the medical field. Cross-domain approaches [2, 29, 54] effectively transfer the knowledge acquired from the training on large-scale datasets (*i.e.*, source domain) to the target domain. Since the actions taken in the target domain may differ from those in the source domain, the time-consuming manual annotation of the action labels is required. Therefore, cross-domain few-shot learning (CD-FSL) [16, 26, 27, 66, 67] emerges as a crucial strategy, merging cross-domain *adaptability* with the efficiency of few-shot learning [7, 25, 47, 49, 55, 65] which only uses a few labeled samples on the target domain. Recently, using unlabeled data on the target domain has proven effective in enhancing the adaptability against the domain without increasing the annotation cost [19, 37, 46]. These approaches further improve the task performance in the CD-FSL setting.

The CD-FSL with unlabeled target data [19,37,46] adopts two meta-training stages (pretraining and domain adaptation) and two meta-testing stages (few-shot training and inference). For example, Dynamic Distillation [19] first trains the visual encoder on the source dataset (pretraining stage), subsequently training with labeled source and unlabeled target data via pseudo distillation [63,64] (domain adaptation stage). Lastly, a classifier is trained on the few-shot support data of novel classes (few-shot training stage) and evaluated on the query samples (inference). Building upon this, CDFSL-V [46] proposed to use the recent self-supervised technique, VideoMAE [52], for the pretraining stage, enabling the pretraining with the unlabeled target data to enhance the adaptability to the target domain on the action recognition task.

Although the CD-FSL task with unlabeled target data has been widely investigated, two significant issues remain when applying the previous works [19, 37, 46] to the egocentric action recognition task.

Adaptability to the Target Domain. Solely relying on visual information is susceptible when performing on the target domain despite using domain adaptation techniques. While the previous works have focused only on RGB modality

for image or video classification tasks, it is known that using multimodal information (*e.g.*, optical flows or poses) helps to mitigate the domain gap between the source and target domains for the action recognition task [33,34]. Especially on the egocentric action recognition task, since the background's visual information vastly changes, solely relying on the RGB information to adapt the domain is infeasible.

Inference Cost. Despite the advancements in the CD-FSL task expanding for videos [46], processing densely-sampled input frames with temporal-aware operations is necessary, making the process computationally expensive [24,57,58]. This computational intensity hampers the practical applications on edge devices with limited resources. Therefore, reducing the inference time is essential for egocentric action recognition.

In this work, for the first time, we study cross-domain few-shot learning for egocentric action recognition using multimodal input and unlabeled target data (Fig. 1). We propose MM-CDFSL, a novel CD-FSL approach utilizing multimodal input and unlabeled target data to enhance the adaptability to the target domain. In our meta-training framework, we first pretrain the domain-adapted and class-discriminative features at the pretraining stage. Subsequently, we conduct multimodal distillation to bridge the domain gap further. This work capitalizes on incorporating multiple modalities to alleviate the extreme domain gaps on the CD-FSL task. Additionally, we consider the computational cost during inference by reducing the number of input tokens without compromising the action recognition accuracy. We aim to simultaneously achieve strong adaptability to the target domain where the shift between the source and target domain is significant while improving the runtime during the inference.

To address the first challenge, we propose to (1) incorporate supervised training on source-labeled data in the pretraining stage and (2) additionally introduce the multimodal distillation stage to transfer the knowledge obtained during the pretraining to a network with RGB input. In the pretraining stage, domain-adapted and class-discriminative features are obtained using labeled source and unlabeled target data. Utilizing VideoMAE [52] for reconstructing source and target inputs, our model aims to capture discriminative and shared representations between source and target domains. This pretraining process is applied independently for all input modalities. Subsequently, we perform multimodal feature distillation on the unlabeled target dataset to transfer knowledge from the teacher models trained on source and target domains for multiple modalities to the student RGB encoder. Since the teacher networks are trained on both source and target data, leveraging only unlabeled target data during multimodal distillation helps improve the adaptability of the student network to the target domain.

To address the challenge of computational efficiency, we propose an *Ensemble Masked Inference*, which reduces the number of input tokens by randomly masking tokens from the input data and then ensembling multiple classification results estimated from these inputs. However, naively masking the input data during inference can lead to a severe performance drop since the feature

extractor and the classifier are trained on the unmasked inputs. Therefore, drawing inspiration from the *Tube Masking* operation in VideoMAE [52], originally proposed to randomly mask input tokens in videos to obtain meaningful representations during pretraining, we propose to apply this operation across all stages, including pretraining, multimodal distillation, few-shot training, and inference. By applying masking throughout these stages, we prevent distribution drift while simultaneously improving inference speed.

In summary, our contributions are as follows:

- We propose a novel, challenging, but practical problem: cross-domain few-shot learning with multimodal input and unlabeled target data in egocentric scenarios.
- We propose MM-CDFSL, a novel approach for the CD-FSL for egocentric action recognition task that utilizes a domain-adapted and class-discriminative pretraining and multimodal feature distillation. Furthermore, we propose ensemble masked inference to reduce the computational cost.
- We simultaneously achieve state-of-the-art performance in terms of the accuracy and inference speed on multiple egocentric action recognition benchmarks [1, 5, 43] with the CD-FSL settings. Specifically, our method outperforms the prior state-of-the-art in accuracy by an average of 6.12/6.10 points for 1-shot/5-shot settings, while our approach is 2.2x faster than the previous approaches.

2   Related Work

First, we discuss the most relevant work on egocentric action recognition and then review previous efforts on cross-domain few-shot learning (CD-FSL). Our work intersects with prior research from several perspectives: few-shot learning, cross-domain adaptability, the utilization of unlabeled target data, and the incorporation of multimodal inputs. To provide a clear and concise comparison, we have summarized the differences in Tab. 1.

Egocentric Action Recognition. Egocentric action recognition has gained popularity [38, 41, 51, 61] with the advent of affordable, lightweight wearable cameras, such as GoPro. This increase in egocentric data has sparked significant interest in understanding and recognizing actions within these videos [36]. One of the key challenges for egocentric action recognition is an environmental bias where egocentric videos are captured in different locations. For example, some researchers [39, 48] have aimed to acquire a domain generalized or domain adapted representation. Also, the lack of large-scale datasets with annotations is a fundamental problem in egocentric vision, especially in the industrial and medical fields. This necessitates the few-shot learning technique, which enables the inference of novel classes with limited labeled samples, thereby fueling research in this domain [11, 60]. In contrast to the previous work, we simultaneously tackle these two key challenges.

On the other hand, incorporating additional multimodal information is essential for identifying activities for egocentric activity recognition. Several stud-

Table 1: Comparative overview of methodological features.

Methods	Few-shot	Cross-domain	Unlabeled target data	Multimodal
[7, 25, 47, 49, 55, 65]	✓			
[14, 28, 53, 59]	✓	✓		
[19, 37, 46]	✓	✓	✓	
[42, 50]				✓
[11]	✓			✓
MM-CDFSL (Ours)	✓	✓	✓	✓

ies have demonstrated that leveraging supplementary modalities at inference time significantly enhances performance [8, 22, 23, 32, 35, 42, 62]. However, acquiring and processing the additional modalities typically demands significant computational resources. This limitation renders these methods cumbersome or impractical, particularly on constrained computing budgets such as embedded devices. To benefit from multimodal information without increasing computational cost at inference time, the distillation technique has attracted numerous researchers [42,50]. Following this success, our approach incorporates multimodal distillation. However, unlike the existing work [42,50], we focus on a challenging but practical scenario: cross-domain and few-shot settings for egocentric action recognition.

Cross-Domain Few-Shot Learning. Guo *et al*. [14] introduced the Cross-Domain Few-Shot Learning (CD-FSL) challenge, which assesses the capability of deep neural networks to adapt image classification to novel domains that present classes not encountered during their training. Research has explored numerous strategies, such as fine-tuning [4], feature-wise transformation [53], prototype learning [59], and the training with multiple domains [28], all aimed at enhancing the networks' generalization capabilities. Among these strategies, using unlabeled target data stands out for its ability to increase adaptability to target domains, offering a practical approach akin to unsupervised domain adaptation [9,15], as the annotation for the target data is not required.

To incorporate unlabeled target data during training, STARTUP [37] and Dynamic Distillation [19] employ a pseudo-labeling technique based on the self-training strategy. Both methods first pretrain the visual feature extractor on the labeled source dataset, subsequently training with student-teacher networks initialized from the pretrained weights. CDFSL-V [46] addresses challenging cross-domain few-shot learning for action recognition. The method incorporates the recent self-supervised technique for video understanding, VideoMAE [52], to utilize unlabeled target data during the pretraining stage, leading to better adaptability to the target domain.

However, neither of these approaches investigates the leverage of multiple modalities, which is crucial in egocentric action recognition, and considers inference time. In contrast, we explore utilizing multiple modalities for adaptability to the target domain and consider the computation cost.
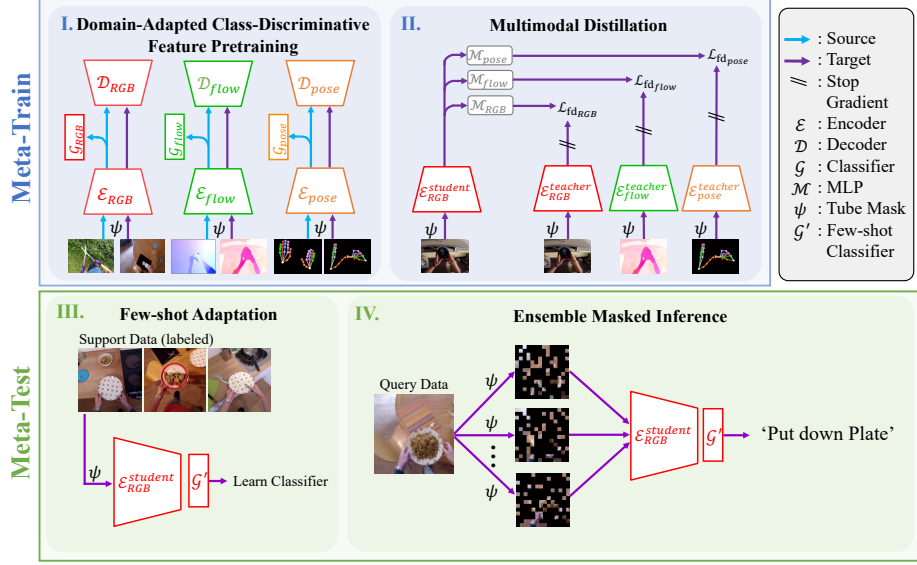
Fig. 2: The framework of our proposed method. Our approach has two meta-training and two meta-testing stages: 1. learning domain-adapted and class-discriminative features for all modalities, 2. distilling the multimodal features into student RGB encoders, 3. few-shot learning for adapting novel classes, and 4. ensemble masked inference using P Tube Masking during inference.

3   Method

The proposed method consists of two meta-training and two meta-testing stages: the first involves domain-adapted and class-discriminative feature pretraining, the second benefits the domain adaptability from multimodal distillation, the third trains a classifier to adapt novel classes using labeled few-shot samples, and the fourth infer the action class of query data while reducing the inference time. We first introduce a new problem setup of the cross-domain few-shot learning task with multimodal input and unlabeled target data (Sec. 3.1). Then, we introduce the proposed method, including (1) pretraining (Sec. 3.2), (2) multimodal distillation (Sec. 3.3), (3) few-shot training and (4) ensemble masked inference (Sec. 3.4). Fig. 2 provides an overview of our approach. For readability, the following section is explained with RGB, optical flow, and heatmap of the hand pose (referred to as *hand pose*) as multimodal information, known as effective modal information for egocentric action recognition. However, our method can use any modality information (*e.g.*, IMU, audio).

3.1   Problem Definition

The goal of the cross-domain few-shot learning (CD-FSL) task with multimodal input and unlabeled target data is to classify novel classes within the target dataset D_T, leveraging both a labeled source dataset D_S and an unlabeled target dataset D_{T_u}. Both D_S and D_{T_u} comprise data across m modalities. Ensure there is no class overlap between the source and target datasets to maintain the integrity of the few-shot learning paradigm. To infer novel classes in the target dataset, the data is split into a support set S, which contains K data from N classes for the N-way K-shot task, and a query set Q consisting of data from only the N classes of the support set, same as the standard few-shot learning setting.

3.2   Domain-adapted and Class Discriminative Feature Pretraining

In the pretraining stage, the VideoMAE [52] for each modal is independently trained to learn the representation shared among the source/target domain and discriminative features on the source domain via two objective functions: (1) joint reconstruction of the source and target domain data and (2) the classification of the action categories on the source dataset. In contrast to the pretraining stage in CDFSL-V [46], which only reconstructs the data on the source and target dataset to mitigate the domain shift, we train the model via a joint optimization of the above two objectives to effectively learn the shared and discriminative feature representation.

Specifically, given an input $\boldsymbol{x}_m \in \mathbb{R}^{T \times H_m \times W_m \times C_m}$ consisting T frames of image size $H_m \times W_m$ with a channel size C_m for modality $m \in \{$RGB, optical flow, hand pose$\}$, respectively, we train a VideoMAE [52], which consists of encoder (feature extractor) $\mathcal{E}_m$ and decoder $\mathcal{D}_m$, per modality m. Following the training framework of VideoMAE, the tokenized input data are randomly masked with the masking ratio ρ_{pretrain} by *Tube Masking* ψ and reconstructed as follows:

$$\hat{\boldsymbol{x}}_m = \mathcal{D}_m(\mathcal{E}_m(\psi(\boldsymbol{x}_m))), \tag{1}$$

where $\hat{\boldsymbol{x}}_m$ denotes reconstructed input. Additionally, a classifier $\mathcal{G}_m$ is employed to input the averaged embedding extracted from $\mathcal{E}_m$ and estimates the logit $l_m \in \mathbb{R}^{n_c}$ (n_c denotes the number of the classes in source dataset) as follows: $l_m = \mathcal{G}_m(\mathcal{E}_m(\psi(\boldsymbol{x}_m)))$.

We train $\mathcal{E}_m$, $\mathcal{D}_m$, and $\mathcal{G}_m$ for each modal with the combination of three losses: the reconstruction loss for source data $\mathcal{L}_{\mathrm{recon}}^{\mathrm{source}}$, the reconstruction loss for target data $\mathcal{L}_{\mathrm{recon}}^{\mathrm{target}}$, and the cross entropy loss for action classification in the source dataset $\mathcal{L}_{\mathrm{ce}}^{\mathrm{source}}$ with a balancing hyperparameter λ_{ce_m}:

$$\mathcal{L}_{\mathrm{pretrain}} = \mathcal{L}_{\mathrm{recon}}^{\mathrm{source}} + \mathcal{L}_{\mathrm{recon}}^{\mathrm{target}} + \lambda_{\mathrm{ce}_m} \mathcal{L}_{\mathrm{ce}}^{\mathrm{source}}. \tag{2}$$

Note that we only calculate $\mathcal{L}_{\mathrm{ce}}^{\mathrm{source}}$ against labeled source data, and we employ L_2 loss for the reconstruction loss.

3.3   Multimodal Distillation

Next, we further improve the adaptability to the target domain and few-shot action recognition accuracy in the cross-domain setting by distilling the domain-adapted multimodal features to the RGB feature extractor $\mathcal{E}_{\mathrm{RGB}}$. Using multiple modalities helps mitigate the domain shift between source and target domain compared to using only RGB modality, as visual information is susceptible to lighting, background, and appearance variations. Incorporating additional modalities like optical flow and hand pose introduces complementary information that is less sensitive to these visual changes. Furthermore, distilling the multimodal features, which are domain-adapted and class-discriminative, into the RGB modality reduces the model complexity while alleviating the domain gap. It aims to imbue the RGB model with domain-adapted characteristics of multimodal learning without the overhead of processing and integrating multiple modality data types during inference.

Given multimodal input for unlabeled target data $\boldsymbol{x}_m^{\mathrm{target}}$, we train the student RGB encoder $\mathcal{E}_{\mathrm{RGB}}^{\mathrm{student}}$ using the teacher encoder $\mathcal{E}_m^{\mathrm{teacher}}$ for all modalities, including RGB. Distilling features from the RGB modality ensures that it regularizes the multimodal distillation from other modalities to prevent forgetting the RGB element and biasing the other modalities. It is noted that all student and teacher encoders are initialized from the pretrained weights in the previous domain-adapted and class-discriminative feature pretraining section, and the weights of teacher encoders are kept frozen during the multimodal distillation stage. Also, we mask the input tokens like the previous self-supervised training, with the same mask ratio ρ_{distill} across the modalities. The masking is adopted as the model is required to make predictions based on masked inputs during inference (See Sec. 3.4). The unlabeled target RGB data is fed to the student RGB encoder; subsequently, the extracted features are projected into modalities by projection layers (*e.g.*, multilayer-perceptrons) $\mathcal{M}_m$:

$$\hat{\boldsymbol{f}}_m = \mathcal{M}_m(\mathcal{E}_{\mathrm{RGB}}^{\mathrm{student}}(\psi(\boldsymbol{x}_{\mathrm{RGB}}^{\mathrm{target}}))), \tag{3}$$

where $\hat{\boldsymbol{f}}_m$ are projected features from RGB to the modality m.

We aim to minimize its L_2 distance from the real embedding of the modality m. Specifically, the loss is computed as a linear combination of the L_2 losses corresponding to each modality, and the feature distillation loss is defined as follows:

$$\mathcal{L}_{\mathrm{fd}_m} = \left\| \mathrm{sg}[\boldsymbol{f}_m] - \hat{\boldsymbol{f}}_m \right\|_2^2, \tag{4}$$

where sg[.] stands for the stop gradient operator that is defined as an identity at forward computation time and has zero partial derivatives, and $\boldsymbol{f}_m$ denotes the extracted features of the unlabeled target data $\boldsymbol{x}_m^{\mathrm{target}}$ from the teacher encoder of the modality m:

$$\boldsymbol{f}_m = \mathcal{E}_m^{\mathrm{teacher}}(\psi(\boldsymbol{x}_m^{\mathrm{target}})). \tag{5}$$

The final training loss in the multimodal distillation stage is the linear combination of these losses for each modality:

$$\mathcal{L}_{\text{distill}} = \sum_m \mathcal{L}_{\text{fd}_m}.$$ (6)

3.4   Ensemble Masked Inference

The computation cost is one of the fundamental problems for a real-time application or inference on limited-resourced devices. Processing all tokens in input frames for the Transformer model for action recognition is computationally expensive [24, 57, 58]; however, the computation cost can be modulated depending on how many tokens from input frames are used. On the one hand, existing methods using the ViT architecture process all tokens from input frames to achieve strong action recognition performance, but this comes at a high computational cost. The attention mechanism demands computational complexity of $\mathcal{O}(I^2)$ where I denotes the number of input tokens. On the other hand, reducing the number of input tokens helps mitigate the computation cost but results in a drop in performance. To this end, we propose the ensemble masked inference to reduce the computation cost by reducing the number of input tokens by masking with the mask ratio ρ_{infer} and alleviate the performance drop by utilizing the ensemble prediction with the ensemble number P: $\mathcal{O}(P((1 - \rho_{\text{infer}})I)^2)$.

Few-Shot Training. Following the existing work [19, 37, 46], we learn a new classifier for adapting the novel classes in the target domain with a few numbers of labeled data. The RGB student encoder is retained, and the classifier head $\mathcal{G}'$ on top of the encoder is trained with a sampled N-way K-shot data from the support set S. The Tube Masking is applied to the input with the mask ratio ρ_{infer} during the few-shot training process, enabling the model to make predictions based on masked input at the inference time. Note that the Tube Mask is varied across the input data from the support set to prevent the model from overly relying on identical masking patterns, which may inadvertently mask all-important content.

Inference. Once the classifier $\mathcal{G}'$ is adapted to the novel classes in the target domain, q samples from each N class from the query set Q are used to evaluate the few-shot action recognition. We apply the Tube Masking ψ with the same mask ratio ρ_{infer} used during the few-shot training. Various mask ratios ρ_{infer} are applied depending on the specific requirements for a trade-off between accuracy and inference speed. Adjusting the mask ratio allows us to tailor the model's performance according to the computational constraints or the precision demands of the application at hand. Furthermore, we adopt ensemble learning to mitigate the drop in performance caused by masking the input frames. We generate P data from one sample by applying varied Tube Masking and then average the predicted probabilities:

$$\hat{y} = \frac{1}{P} \sum_p \text{Softmax}(\mathcal{G}'(\mathcal{E}_{\text{RGB}}(\psi(\boldsymbol{x}_{\text{RGB}})))).$$ (7)

4   Experiments

4.1   Datasets

We employ the most large-scale egocentric video dataset, Ego4D [12] as a source dataset and multiple egocentric datasets [1, 5, 43] as a target dataset.

Ego4D [12]. The Ego4D dataset is one of the large-scale egocentric video datasets in a daily life domain. It contains 3,670 hours of egocentric videos of people performing diverse tasks, such as gardening or crafting, and is collected by 931 people from 74 locations across nine different countries worldwide. We utilize annotations for short-term action anticipation, including both the clip frame and the corresponding time to contact for the egocentric action recognition task. For our experiment, we constructed annotations by selecting clip frames, including the frame at which contact occurs within the input sequence. This results in a 204-class action recognition dataset with 15.5k training clips.

EPIC-Kitchens [5]. EPIC-Kitchens is a dataset captured in the cooking domain, where 32 participants have recorded 432 egocentric videos. The dataset encompasses a total of 286 action classes. For our cross-domain few-shot study, we partitioned these videos into training (unlabeled) and validation subsets, comprising 58 and 228 action classes with 3.6k and 17.2k video clips, respectively. To adhere to the few-shot learning task settings, we ensured no class overlap between the training (unlabeled) and validation sets.

MECCANO [43]. The MECCANO dataset was collected in an industrial-like setting, where 20 participants were asked to assemble a toy model of a motorbike. We partitioned the MECCANO dataset into training and validation subsets. This partition yielded 10 and 40 action classes, accompanied by 1.3k and 7.4k action segments for the training and validation sets.

WEAR [1]. The WEAR dataset, curated explicitly for human activity recognition within the outdoor sports domain, features egocentric video data capturing a wide array of athletic activities in natural settings. It encompasses recordings from 18 participants engaging in 18 distinct workout activities across ten diverse outdoor locations. To facilitate our study, we divided the dataset into training and validation subsets, allocating 3 and 15 activity classes to each, accompanied by 1.8k and 8.0k activity segments, respectively.

4.2   Implementation Details

Experimental Setup. We utilize the three modalities: RGB, optical flow, and hand pose. For each modality, we select an input sequence comprising $T = 16$ frames sampled at a frequency of 8 FPS (frames per second). Spatial dimensions are standardized at 224×224 for RGB and optical flow inputs, while hand pose inputs, represented by the heatmap, are resized to 56×56. The number of channels C_m are 3, 2, and 21 for RGB, optical flow, and hand pose, respectively. We employ FlowFormer [18] for estimating the optical flow between consecutive frames. For the prediction of 2D hand keypoints, which include 21 joints, we use RTMPose [20] trained on five public hand pose datasets available through

MMPose [3]. Subsequently, the Gaussian heatmap of size 56×56 is produced, with a standard deviation σ set to 10. Following the VideoMAE [52] experimental setup, we adopt the mask ratio ρ_{pretrain} of 0.9 during our pretraining stage. In addition, we use the mask ratio $\rho_{\mathrm{distill}} = 0.75$ during the multimodal distillation for all experiments. We use a machine equipped with Intel Xeon W-3235 CPU, 128GB RAM, and the NVIDIA Titan RTX GPU to compute the inference speed.

Training. In the pretraining stage, we train the model, which consists of ViT-S models, pretrained on the Kinetics-400 dataset [21] and a classifier head, which is attached to the ViT-S backbone, for all modalities for 100 epochs. For training settings, we generally follow the VideoMAE [52]. During the multimodal distillation stage, we train the student RGB model for 100 epochs using the AdamW optimizer [31], with a peak learning rate of $2e - 3$, linearly increased for the first 10 epochs of the training and decreased to $1e - 6$ until the end of training with cosine decay [30]. Note that we linearly scaled the peak learning rate with respect to the overall batch size. Regarding the parameters for the loss function, we empirically adapt the balancing hyperparameters λ_{ce_m} in Eq. (2) to $5e - 2$ for RGB, and $1e - 2$ for optical flow and hand pose input modality.

Evaluation Metrics. Following the existing CD-FSL work, we report the top-1 accuracy on the query set Q in the target validation set over 600 runs to measure action recognition performance. To benchmark efficiency, we quantify the model's performance by measuring the forward pass time during inference. We report the inference time averaged over 600 iterations.

4.3   Comparison Methods

- **Random Initialization** is used as our baseline for the experiment. This method entails learning a logistic regression classifier on top of an untrained VideoMAE encoder, providing a fundamental comparison point for evaluating the efficacy of more sophisticated training strategies.
- **STARTUP++** [37] is a modified version of STARTUP, used in CDFSL-V for a fair comparison. It replaces supervised training during the pretraining stage of the STARTUP with self-supervised pretraining on the source dataset using VideoMAE. Subsequently, following STARTUP, it performs the representation learning with the KL-divergence loss and self-supervised contrastive loss [3].
- **Dynamic Distillation++** [19] adopts self-supervised training instead of supervised training, similar to the STARTUP++, to ensure a fair comparison. Then, it trains the student network with pseudo-labeling while dynamically updating the teacher network's parameters throughout training. This approach involves aligning the student network's predictions for strongly augmented versions of unlabeled target data with the teacher network's weakly augmented counterpart.
- **CDFSL-V** [46] proposed to use the recent self-supervised technique, Video-MAE [52], for the pretraining stage, enabling the pretraining with the unla-

---

[3] https://github.com/open-mmlab/mmpose

Table 2: Cross-domain few-shot action recognition accuracy and inference time. We assess the performance of 5-way 1-shot and 5-shot top-1 action recognition accuracy on three egocentric datasets, EPIC-Kitchens (EPIC), MECCANO (MEC), WEAR, and the average across these datasets (Avg.). The Ego4D dataset is used as the source dataset. We report an average of 600 runs of few-shot evaluation. We present the results of our method that adopts the mask ratio $\rho_{\text{infer}} = 0.75$ and ensemble number $P = 2$. Note that existing methods use all tokens from input frames; thus, they have the same value as inference time. The best values are shown in **bold**.

Method	Runtime (ms)	GFLOPs	Memory (MiB)	1-shot				5-shot			
				EPIC	MEC	WEAR	Avg.	EPIC	MEC	WEAR	Avg.
Random Initialization				29.20	23.10	25.96	26.09	40.31	27.04	38.71	35.35
STARTUP++ [37]	22.1	68.5	2782	35.17	26.85	39.15	33.73	50.23	34.05	59.88	48.05
Dynamic Distillation++ [19]				36.96	27.87	35.84	33.56	53.78	**37.87**	56.23	49.29
CDFSL-V [46]				38.17	25.94	39.10	34.40	53.72	35.79	58.28	49.26
Ours	**9.64**	**37.0**	**968**	**41.97**	**28.34**	**51.25**	**40.52**	**58.70**	37.80	**69.57**	**55.36**

> beled target data to enhance the adaptability to the target domain. CDFSL-V also adopts pseudo distillation as a consistency regularization, similar to Dynamic Distillation, during the second training stage.

Note that for all comparison methods except random initialization, the weights are initialized from the pretrained model on the Kinetics-400 at the pretraining stage, the same as our initialization.

4.4   Quantitative Comparison

We compare the performance of 5-way 1-shot and 5-shot top-1 action recognition accuracy and inference time with the prior methods, trained on the Ego4D [12] dataset as the source domain, on three egocentric datasets as the target domain: EPIC-Kitchens [5] in the cooking domain, MECCANO [43] in the industrial-like domain, and WEAR [1] in the outdoor workout domain. We report the accuracy and inference speed of our proposed method with the mask ratio of ρ_{infer} of 0.75, using the ensemble number $P = 2$. Tab. 2 shows that the proposed method consistently outperforms the state-of-the-art CD-FSL with unlabeled target data methods regarding action recognition accuracy and inference speed.

Few-Shot Action Recognition Accuracy. Our proposed model outperforms the CDFSL-V by 6.10 points and the Dynamic Distillation++ by 6.07 points on the average of three datasets regarding the 5-shot action recognition accuracy. Significant improvement can be seen in 5-shot accuracy on the WEAR dataset (from 58.28 to 69.57) as our method leverages the multiple modalities during training, further mitigating the domain gap between source and target. Also, our model outperforms the CDFSL-V by 6.12 points and the Dynamic Distillation++ by 6.95 points on average of the 1-shot setting, even though only one labeled target sample from the support set S is available during the few-shot training stage.

Table 3: Loss component ablation study in the pretraining stage. *Only reconstruction* is when $\lambda_{ce_m} = 0$ in Eq. (2).

Method	1-shot	5-shot
Only reconstruction	35.42	49.82
Ours	**41.97**	**58.70**

Table 4: Ablation study on the multimodal distillation stage. *Only RGB Training* row shows the accuracy without distilling the multimodal information at the second training stage.

Method	1-shot	5-shot
Only RGB Training	46.17	67.19
RGB+Pose	49.39	67.90
Ours	**51.25**	**69.57**

Inference Time. From Tab. 2, our approach achieved 2.2x faster inference speed than the previous approaches while achieving state-of-the-art few-shot classification performance. The approaches in Tab. 2 employ the same architecture (ViT-S) and the same number of input tokens; thus, their inference times remain consistent across implementations. Our approach enhances efficiency and retains high accuracy, making it particularly advantageous for applications demanding online processing without compromising prediction accuracy.

4.5 Ablation Analysis

In this section, we conduct the ablation study of the proposed framework.

Class-discriminative Training in the Pretraining Stage. We conduct an ablation analysis on employing the cross-entropy loss during the pretraining stage in Eq. (2) to assess its impact on class-discriminative feature learning in conjunction with self-supervised learning. We report the performance of the best trade-off model, which incorporates multimodal feature distillation to determine the impact of this cross-entropy loss ablation. As shown in Tab. 3, the results verify the importance of cross-entropy loss in obtaining class-discriminative representations, thereby significantly boosting the model's accuracy and reliability for few-shot action recognition in cross-domain scenarios.

Multimodal Distillation. We also conduct an ablation study during the second stage of training to assess the benefits of multimodal feature distillation. Tab. 4 presents the 1-shot and 5-shot few-shot action recognition accuracy on the WEAR dataset. The results confirm that omitting the multimodal distillation stage leads to a decrease in action recognition accuracy. This decline underscores the significance of multimodal distillation for enhancing performance in cross-domain few-shot action recognition tasks.

Accuracy and Speed Trade-off. The ablation study examines the impact of the masking ratio ρ_{infer}, applied to input frames during inference, and the ensemble number P. This investigation aims to explore the trade-off between model accuracy and inference speed, providing insights into how variations in ρ_{infer} and P influence the overall performance and efficiency of the model.

Fig. 3 illustrates the accuracy vs. inference time trade-off on the EPIC-Kitchens dataset. Our most accurate model ($\rho_{\text{infer}} = 0, P = 1$) significantly surpasses CDFSL-V by 5.23 points with the same inference speed as the existing method. Moreover, our fastest model ($\rho_{\text{infer}} = 0.9, P = 1$) achieves 7.65 ms in inference speed, which is 2.9x faster than the previous method and still outperforms the previous method in terms of accuracy. Although the fastest model suffers a drop in performance compared to our most accurate model, this performance degradation is modulated by increasing the number of ensemble predictions without sacrificing the inference speed (*e.g.* $P = 3$).
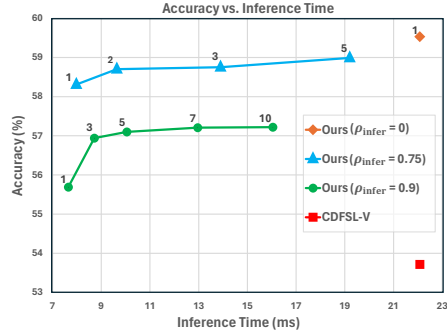

Fig. 3: **Accuracy vs. inference time**. The trade-off analysis between action recognition accuracy and inference speed is conducted for the existing method and our proposed approach, examining various masking ratios ρ_{infer} and ensemble numbers P. The number near the plots for our proposed method denotes the ensemble number.

5 Conclusion

Conclusion. We present MM-CDFSL, the first work to explore the multimodal data for egocentric action recognition in cross-domain and few-shot settings. We propose training the models for each input modality during the first pretraining stage to gain domain-adapted and class-discriminative features. Then, we perform multimodal distillation to the student RGB models using teacher models for all modalities to mitigate the domain gap further. Moreover, we propose ensemble masked inference to reduce the computation cost during inference by masking the input frames while alleviating the drop in performance via ensemble learning. Experiments on egocentric datasets from three domains, EPIC-Kitchens, MEC-CANO, and WEAR datasets, demonstrate that our approach outperforms the state-of-the-art CD-FSL with unlabeled target data methods regarding action recognition accuracy and inference speed.

Limitations and future work. Our proposed method leverages the optical flow and 2D hand keypoints based on the off-the-shelf optical flow estimator [18] and 2D hand keypoints detector [20]. Thus, the bias and errors from the off-the-shelf detector may still affect the input modality information. In addition, our approach applies constant loss weights for feature distillation losses during the multimodal distillation process, regardless of the specific target dataset. This strategy does not account for the varying significance of modalities based on the target domain; for instance, motion information may hold greater relevance than hand pose data in outdoor environments. Dynamically adjusting distillation weights according to the modality's relevance in the target domain is crucial for achieving more targeted and efficient training outcomes. We will leave this for our future efforts.

References

1. Bock, M., Moeller, M., Van Laerhoven, K., Kuehne, H.: Wear: A multimodal dataset for wearable and egocentric video activity recognition. arXiv preprint arXiv:2304.05088 (2023) 4, 10, 12, 1

2. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: CVPR (2017) 2

3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020) 11

4. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C., Huang, J.B.: A closer look at few-shot classification. In: ICLR (2019) 5

5. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018) 4, 10, 12, 1

6. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjupalli, D., Frost, D., Miller, E., Mueggler, E., Oleinik, E., Zhang, F., Somasundaram, G., Solaira, G., Lanaras, H., Howard-Jenkins, H., Tang, H., Kim, H.J., Rivera, J., Luo, J., Dong, J., Straub, J., Bailey, K., Eckenhoff, K., Ma, L., Pesqueira, L., Schwesinger, M., Monge, M., Yang, N., Charron, N., Raina, N., Parkhi, O., Borschowa, P., Moulon, P., Gupta, P., Mur-Artal, R., Pennington, R., Kulkarni, S., Miglani, S., Gondi, S., Solanki, S., Diener, S., Cheng, S., Green, S., Saarinen, S., Patra, S., Mourikis, T., Whelan, T., Singh, T., Balntas, V., Baiyya, V., Dreewes, W., Pan, X., Lou, Y., Zhao, Y., Mansour, Y., Zou, Y., Lv, Z., Wang, Z., Yan, M., Ren, C., Nardi, R.D., Newcombe, R.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023) 1

7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017) 2, 5

8. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal Transformer for Video Retrieval. In: European Conference on Computer Vision (ECCV) (2020) 5

9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015) 5

10. Gong, X., Mohan, S., Dhingra, N., Bazin, J.C., Li, Y., Wang, Z., Ranjan, R.: Mmg-ego4d: Multimodal generalization in egocentric action recognition. In: CVPR (2023) 1

11. Gong, X., Mohan, S., Dhingra, N., Bazin, J.C., Li, Y., Wang, Z., Ranjan, R.: Mmg-ego4d: Multimodal generalization in egocentric action recognition. In: CVPR (2023) 4, 5

12. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Ruiz, P., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T.,

Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022) 1, 10, 12

13. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. arXiv preprint arXiv:2311.18259 (2023) 1

14. Guo, Y., Codella, N.C., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: ECCV (2020) 5

15. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML (2018) 5

16. Hu, Y., Ma, A.J.: Adversarial feature augmentation for cross-domain few-shot classification. In: ECCV (2022) 2

17. Huang, C., Tian, Y., Kumar, A., Xu, C.: Egocentric audio-visual object localization. In: CVPR (2023) 1

18. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: FlowFormer: A transformer architecture for optical flow. In: ECCV (2022) 10, 14

19. Islam, A., Chen, C.F., Panda, R., Karlinsky, L., Feris, R., Radke, R.: Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. In: NeurIPS (2021) 2, 5, 9, 11, 12

20. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtmpose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399 (2023) 10, 14

21. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 11

22. Kazakos, E., Huh, J., Nagrani, A., Zisserman, A., Damen, D.: With a little help from my temporal context: Multimodal egocentric action recognition. In: BMVC (2021) 5

23. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (2019) 5

24. Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., Gong, B.: Movinets: Mobile video networks for efficient video recognition. In: CVPR (2021) 3, 9

25. Lee, K., Maji, S., Ravichandran, A., Soatto, S.: Meta-learning with differentiable convex optimization. In: CVPR (2019) 2, 5
26. Li, W.H., Liu, X., Bilen, H.: Cross-domain few-shot learning with task-specific adapters. In: CVPR (2022) 2
27. Liang, H., Zhang, Q., Dai, P., Lu, J.: Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In: ICCV (2021) 2
28. Liu, Y., Lee, J., Zhu, L., Chen, L., Shi, H., Yang, Y.: A multi-mode modulator for multi-domain few-shot classification. In: ICCV (2021) 5
29. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation (2018) 2
30. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: ICLR (2017) 11
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 11
32. Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., Darrell, T.: Something-Else: Compositional Action Recognition With Spatial-Temporal Interaction Networks. In: CVPR (2020) 5
33. Moon, G., Kwon, H., Lee, K.M., Cho, M.: Integralaction: Pose-driven feature integration for robust human action recognition in videos. In: CVPR (2021) 3
34. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR (2020) 3
35. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. In: NeurIPS (2021) 5
36. Núñez-Marcos, A., Azkune, G., Arganda-Carreras, I.: Egocentric vision-based action recognition: A survey. Neurocomputing **472**, 175–197 (2022) 4
37. Phoo, C.P., Hariharan, B.: Self-training for few-shot transfer across extreme task differences. In: ICLR (2021) 2, 5, 9, 11, 12
38. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR (2012) 4
39. Plizzari, C., Perrett, T., Caputo, B., Damen, D.: What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. In: ICCV (2023) 4
40. Plizzari, C., Planamente, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., Caputo, B.: E2(go)motion: Motion augmented event stream for egocentric action recognition. In: CVPR (2022) 1
41. Poleg, Y., Ephrat, A., Peleg, S., Arora, C.: Compact cnn for indexing egocentric videos. In: WACV (2016) 4
42. Radevski, G., Grujicic, D., Blaschko, M., Moens, M.F., Tuytelaars, T.: Multimodal distillation for egocentric action recognition. In: ICCV (2023) 5
43. Ragusa, F., Furnari, A., Farinella, G.M.: Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. Computer Vision and Image Understanding (CVIU) (2023) 4, 10, 12, 1
44. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: Spotem: Efficient video search for episodic memory. In: ICML (2023) 1
45. Ryan, F., Jiang, H., Shukla, A., Rehg, J.M., Ithapu, V.K.: Egocentric auditory attention localization in conversations. In: CVPR (2023) 1
46. Samarasinghe, S., Rizve, M.N., Kardan, N., Shah, M.: Cdfsl-v: Cross-domain few-shot learning for videos. In: ICCV (2023) 2, 3, 5, 7, 9, 11, 12
47. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017) 2, 5

48. Song, X., Zhao, S., Yang, J., Yue, H., Xu, P., Hu, R., Chai, H.: Spatio-temporal contrastive domain adaptation for action recognition. In: CVPR (2021) 4
49. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR (2018) 2, 5
50. Tan, S., Nagarajan, T., Grauman, K.: Egodistill: Egocentric head motion distillation for efficient video understanding. In: NeurIPS (2023) 5
51. Tekin, B., Bogo, F., Pollefeys, M.: H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In: CVPR (2019) 4
52. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: NeurIPS (2022) 2, 3, 4, 5, 7, 11
53. Tseng, H.Y., Lee, H.Y., Huang, J.B., Yang, M.H.: Cross-domain few-shot classification via learned feature-wise transformation. In: ICLR (2020) 5
54. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017) 2
55. Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., Wierstra, D.: Matching networks for one shot learning. In: NeurIPS (2016) 2, 5
56. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: CVPR (2023) 1
57. Wang, J., Yang, X., Li, H., Liu, L., Wu, Z., Jiang, Y.G.: Efficient video transformers with spatial-temporal token selection. In: ECCV (2022) 3, 9
58. Wang, J., Sun, Z., Qian, Y., Gong, D., Sun, X., Lin, M., Pagnucco, M., Song, Y.: Maximizing spatio-temporal entropy of deep 3d CNNs for efficient video recognition. In: ICLR (2023) 3, 9
59. Wang, W., Duan, L., Wang, Y., En, Q., Fan, J., Zhang, Z.: Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer. In: CVPR (2022) 5
60. Wang, X., Zhang, S., Qing, Z., Tang, M., Zuo, Z., Gao, C., Jin, R., Sang, N.: Hybrid relation guided set matching for few-shot action recognition. In: CVPR. pp. 19948–19957 (2022) 4
61. Wang, X., Zhu, L., Wang, H., Yang, Y.: Interactive prototype learning for egocentric action recognition. In: ICCV (2021) 4
62. Wang, X., Gupta, A.: Videos as space-time region graphs. In: ECCV (2018) 5
63. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR (2020) 2
64. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: CVPR (2021) 2
65. ZHANG, R., Che, T., Ghahramani, Z., Bengio, Y., Song, Y.: Metagan: An adversarial approach to few-shot learning. In: NeurIPS (2018) 2, 5
66. Zhao, A., Ding, M., Lu, Z., Xiang, T., Niu, Y., Guan, J., Wen, J.R.: Domain-adaptive few-shot learning. In: WACV (2021) 2
67. Zhou, F., Wang, P., Zhang, L., Wei, W., Zhang, Y.: Revisiting prototypical network for cross domain few-shot learning. In: CVPR (2023) 2