In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021

Ego-Exo: Transferring Visual Representations from Third-person to First-person Videos

Yanghao Li[1] Tushar Nagarajan[1,2] Bo Xiong[1] Kristen Grauman[1,2]
[1] Facebook AI Research [2] UT Austin
lyttonhao@fb.com, tushar@cs.utexas.edu, bxiong@fb.com, grauman@fb.com

Abstract

We introduce an approach for pre-training egocentric *video models using large-scale* third-person *video datasets. Learning from purely egocentric data is limited by low dataset scale and diversity, while using purely exocentric (third-person) data introduces a large domain mismatch. Our idea is to discover latent signals in third-person video that are predictive of key egocentric-specific properties. Incorporating these signals as knowledge distillation losses during pre-training results in models that benefit from both the scale and diversity of third-person video data, as well as representations that capture salient egocentric properties. Our experiments show that our "Ego-Exo" framework can be seamlessly integrated into standard video models; it outperforms all baselines when fine-tuned for egocentric activity recognition, achieving state-of-the-art results on Charades-Ego and EPIC-Kitchens-100.*

1. Introduction

Egocentric video captured by wearable cameras offers a unique perspective into human behavior. It is the subject of a recent surge in research interest in first-person activity recognition [35, 77], anticipation [21, 1], and video summarization [36, 74, 13] with many valuable future applications in augmented reality and robotics. Compared to third-person videos, egocentric videos show the world through a distinct viewpoint, encode characteristic egocentric motion patterns due to body and head movements, and have a unique focus on hands, objects, and faces, driven by the camera wearer's attention and interaction with their surroundings.

However, these unique properties also present a fundamental challenge for video understanding. On the one hand, learning models purely from egocentric data are limited by dataset scale. Current egocentric video datasets are small (e.g., 90k clips in EPIC-Kitchens-100 [11] vs. 650k in Kinetics-700 [34]) and lack diversity (e.g., videos only in kitchen scenes). On the other hand, a purely exocentric approach that uses more readily available third-person video—the status-quo for pre-training video models [19, 67, 76, 69]—ignores the unique properties of egocentric video and faces a major domain mismatch. Prior work has shown that this latter strategy, though popular, is insufficient: pre-training egocentric action recognition models with third-person data alone produces significantly worse results than pre-training with first-person data [62]. In an attempt to bridge the domain gap, prior work explores traditional embedding learning [61, 75] or domain adaptation approaches [10], but they require *paired* egocentric and third-person videos that are either concurrently recorded or annotated for the same set of activities, which are difficult to collect and hence severely limit their scope.

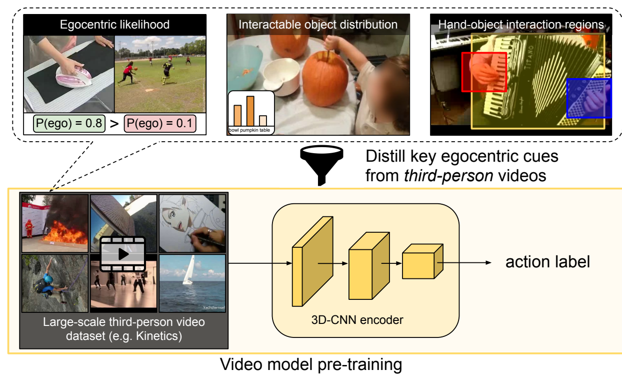

Figure 1: **Main idea.** We extract key egocentric signals from large-scale third-person data and distill them into the video backbone during pre-training to guide feature learning for egocentric video tasks with wearable camera video.

Despite their differences, we hypothesize that the exocentric view of activity should in fact inform the egocentric view. First, humans are able to watch videos of other people performing activities and map actions into their own (egocentric) perspective; babies in part learn new skills in just this manner [48, 55]. Second, exocentric video is not devoid of person-centered cues. For example, a close-up instructional video captured from the third-person view may nonetheless highlight substantial hand-object interactions; or video captured with a hand-held phone may follow an event (e.g., a parade) as it unfolds with attentional cues related to a head-mounted camera.

Building on this premise, in this work we ask: *"How can we best utilize current video datasets to pre-train egocentric video models?"* Our key idea is to discover latent signals in third-person video that approximate egocentric-specific properties. To that end, we introduce a feature learning approach in which ego-video features are guided by both exo-video activity labels and (unlabeled) ego-video cues, to better align traditional third-person video pre-training with downstream egocentric video tasks.

Specifically, we introduce a series of ego-inspired tasks that require the video model to be predictive of manipulated objects, spatiotemporal hand-object interaction regions, and general egocentricity. Then we incorporate these tasks into training as knowledge-distillation losses to supplement an action classification pre-training objective on third-person video. See Fig. 1.

By design, our video models can continue to enjoy large amounts of labeled third-person training data, while simultaneously embedding egocentric signals into the learned features, making them a suitable drop-in replacement for traditional video encoders for egocentric video tasks. Finally, our approach does not assume any paired or activity-labeled egocentric videos during pre-training; the egocentric signals are directly inferred from third-person video.

Our experiments on three challenging egocentric video datasets show that our "Ego-Exo" framework learns strong egocentric feature representations from third-person video. On Charades-Ego [62], our model improves over models pre-trained on Kinetics—the standard pre-training and fine-tuning paradigm—by +3.26 mAP, and outperforms methods that specifically aim to bridge the domain gap between viewpoints. Finally, our pre-trained model achieves state-of-the-art results on EPIC-Kitchens-100 [11], the largest available first-person dataset.

2. Related Work

Egocentric video understanding The unique viewpoint in egocentric video presents interesting research challenges including action recognition and anticipation [77, 21, 59, 1], daily life summary generation [36, 74], inferring body pose [32, 51], and estimating gaze [37, 31]. Several egocentric video datasets have been created to support these challenges [12, 38, 54, 62]. Model architectures proposed for these tasks include multi-stream networks [46, 38, 35, 68], recurrent networks [22, 21, 65], 3D conv nets [53, 44] and spatially grounded topological graph models [50].

These architectures vary significantly, but all use video encoders that are similarly pre-trained with third-person video datasets, despite being applied to egocentric video tasks. In contrast, we introduce key egocentric losses during *exocentric video* pre-training that bridge the domain gap when applied to downstream egocentric video tasks.

Joint first/third person video understanding Several strategies have been proposed to address the domain gap between first and third person video. Prior work learns *viewpoint-invariant* representations using embedding learning methods, and applies them to action recognition [63, 61, 3], video summarization [30], image retrieval [17], person segmentation [72], and attention-driven gaze prediction [75]. Image generation methods [16, 56, 57, 41] use generative adversarial frameworks to synthesize one viewpoint from the other. Viewpoint invariance has also been treated as a *domain adaptation* task in prior work, adapting third-person video models for overhead drone-footage [10]. Other methods use egocentric video as a modality to supplement top-view footage to improve identification and tracking models [2, 4, 5, 73].

The above methods require *paired* datasets that are either simultaneously recorded or that share the same labels for instances across viewpoints. In contrast, our method leverages only third-person video datasets, but is augmented with pseudo-labels (derived from first-person models) to learn egocentric video representations, thus circumventing both the need for first-person video during pre-training and the need for paired labeled data.

Knowledge distillation for video In knowledge distillation (KD), one network is trained to reproduce the outputs of another [29]. Distillation serves to compress models [29, 49, 8] or incorporate privileged information from alternate tasks [43]. In videos, KD can incorporate information from alternate modalities like audio [24, 6], depth and flow [28, 64], or a combination [45], and object-level information [52]. In the context of self-supervised learning, prior work assigns weak image labels to video instances as supervision for video-level models [25]. In contrast, we use inferred weak labels that are relevant to the egocentric domain, and we use them alongside third-person video labels, rather than in place of them during pre-training.

Egocentric cues in video understanding models Egocentric video offers several unique cues that have been leveraged to improve video understanding models [39, 46]. These include attention mechanisms from gaze and motor attention [47, 38, 42], active object detection [20, 7, 15, 70], and hands in contact [66, 60, 33]. We are also interested in such important egocentric cues, but unlike prior work we do not train models on labeled egocentric video datasets to detect them. Instead, we embed labels predicted for these cues as auxiliary losses in *third-person video* models to steer feature learning towards egocentric relevant features. Unlike any of these prior methods, our goal is to leverage third-person video to pre-train first-person video models.

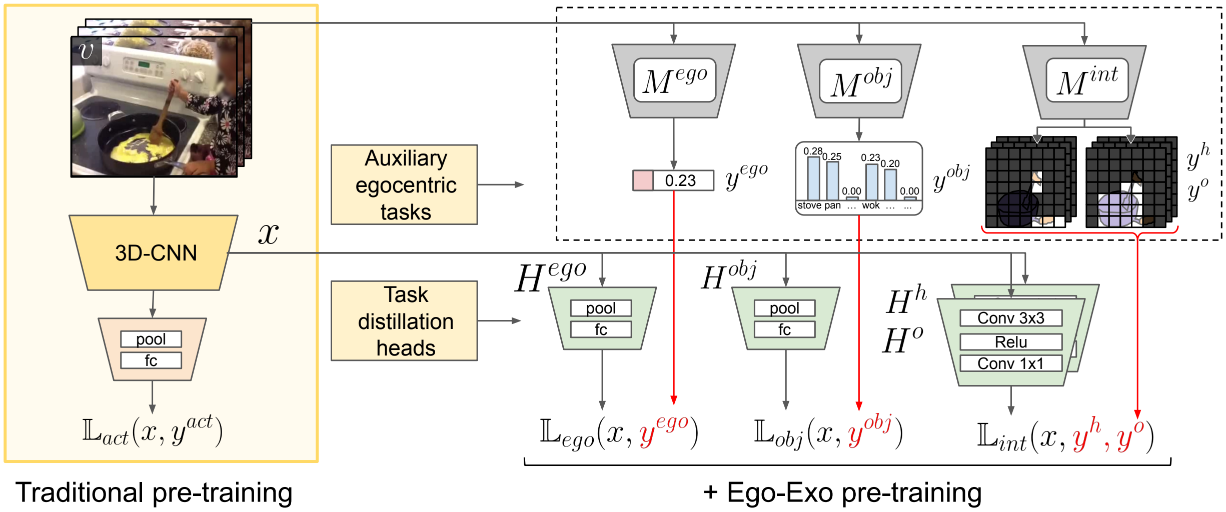

Figure 2: **Ego-Exo framework**. To enhance traditional pre-training (left panel), we generate soft-labels for *third-person videos* from a set of pre-trained egocentric models (top-right) that capture a variety of key egocentric signals (Sec 3.3), and we train distillation modules to approximate the responses of these models (bottom-right). Once pre-trained, the video backbone can be directly fine-tuned for a downstream egocentric task.

3. Ego-Exo Approach

Our goal is to learn egocentric video representations from third-person video datasets, by discovering and distilling important cues about hands, objects, and interactions (albeit in a different viewpoint) that are relevant to egocentric activity during pre-training. To do this, we automatically assign third-person video instances with various egocentric pseudo-labels that span simple similarity scores to complex spatiotemporal attention maps, and then we introduce auxiliary losses that force our features to be predictive of these pseudo-labels. On the one hand, our approach retains the benefits of large-scale third-person video and the original action classification task to guide general video feature learning. On the other hand, we steer feature learning towards better egocentric features using automatically generated egocentric labels, as opposed to collecting manually labeled instances.

In the following sections, we first describe the traditional video pretraining framework (Sec 3.1) and how we incorporate our auxiliary loss terms into it (Sec 3.2). Next we describe the three egocentric tasks we use, namely Ego-Score, Object-Score, and Interaction-Map (Sec 3.3). Finally, we present our full training and evaluation pipeline in Sec 3.4.

3.1. Video model pre-training

Video models benefit greatly from strong initializations. The standard procedure for training egocentric video models is thus to first pre-train models using large-scale third-person video datasets, and then fine-tune for a specific downstream task.

More formally, we are provided with a large-scale third-person (exocentric) video dataset $\mathcal{V}_{exo}$. In pre-training, each video instance $v \in \mathcal{V}_{exo}$ consists of T frames $\{f_1, ..., f_T\}$ and an associated action label y^{act}. These frames are encoded into a series of N spatiotemporal clip features $\{x_1, ..., x_N\}$, where $x_i \in \mathbb{R}^{c\times t\times h\times w}$, using a video encoder *backbone* (e.g., a 3D CNN model). These features are then passed to a classifier *head*, which spatiotemporally pools the feature and uses a linear classifier to generate the predicted action class $\hat{y}^{act}$. Predictions are typically generated for each clip and then averaged to generate video-level predictions. The network is trained to minimize the cross entropy loss $\mathbb{L}_{act}(y^{act}, \hat{y}^{act})$. See Fig 2 (left panel).

Once pre-trained, the backbone weights are retained, the head is replaced with a task-specific classifier, and the new network is trained with instances from a target egocentric dataset $\mathcal{V}_{ego}$ to predict egocentric video labels.

3.2. Ego-Exo pre-training

Third-person pre-training alone results in strong, general-purpose video features. However, it ignores important egocentric signals and introduces a domain gap that limits its utility for downstream egocentric tasks. We introduce auxiliary egocentric task losses to overcome this gap.

Specifically, along with datasets $\mathcal{V}_{exo}$ and $\mathcal{V}_{ego}$, we assume access to off-the-shelf video models that address a set of egocentric video understanding tasks. For each task τ, the model M^{τ} takes as input a video (as either frames or clips) and generates predicted labels y^{τ}. We use these pre-trained models to associate egocentric *pseudo-labels* to the third-person video instances in $\mathcal{V}_{exo}$. We stress that the videos in $\mathcal{V}_{exo}$ are *not* manually labeled for any task τ.

We introduce a task-specific head H^{τ} for each task that is trained to approximate these pseudo-labels for each video instance, leading to an auxiliary loss term $\mathbb{L}_\tau(H^\tau(v), y^\tau)$. Each head is trained to approximate the response of an egocentric video model when applied to a third-person video instance, and thus can be seen as a knowledge-distillation mechanism that distills information from the egocentric tasks into the video encoder model. The final pre-training objective is the combination of the action classification loss $\mathbb{L}_{act}$ and each of the auxiliary loss terms $\mathbb{L}_\tau$. See Fig 2 for our full framework.

Note that these pseudo-labels vary in structure and semantics, ranging from scalar scores (e.g., to characterize how *egocentric-like* a third-person video is), categorical labels (e.g., to identify the manipulated objects in video) and spatiotemporal attention maps (e.g., to characterize hand-object interaction regions). Moreover, these labels are egocentric-specific, but they are *automatically* generated for third-person video instances. This diverse combination leads to robust feature learning for egocentric video, as our experiments will show. Once pre-trained, we can retain our *enhanced* backbone weights to fine-tune on an egocentric video task using data from $\mathcal{V}_{ego}$.

3.3. Auxiliary egocentric tasks

Next, we describe each task we use, how we source M^τ and pseudo-labels y^τ, the loss terms $\mathbb{L}_\tau$, and their relevance to egocentric feature learning. Note that no egocentric activity labels are used for the task models, and each task model is applied to third-person video instances in $\mathcal{V}_{exo}$.

Ego-Score: Discriminating ego videos. A good egocentric video representation should be able to capture the underlying differences between first- and third-person videos, to discriminate between the two viewpoints. Based on this motivation, we design an *Ego-Score* task τ^{ego} to characterize the egocentricity likelihood of the video.

For this, we train a binary *ego-classifier* M^{ego} on the Charades-Ego dataset [62], which has both egocentric and third-person videos of indoor activities involving object interactions. While the dataset offers paired instances showing the same activity from two views, our method does not use this pairing information or egocentric activity labels. It uses only the binary labels indicating if a sample is egocentric or exocentric. Please see Supp. for more training details and an ablation study about the pairing information.

We use this trained classifier to estimate the real-valued pseudo task-labels y^{ego} for each video in our pre-training framework described in Sec 3.2. We sample multiple clips from the same video and average their score to generate a video-level label. Formally, for a video v with N clips $\{x_1, ..., x_N\}$ we generate scores:

$$y_i^{ego}(v) = \frac{\exp(\frac{1}{N\beta}\sum_n z_i^{ego}(x_n))}{\sum_j \exp(\frac{1}{N\beta}\sum_n z_j^{ego}(x_n))}, \tag{1}$$

where β is a scalar temperature parameter, $z_i^{ego}(x_n)$ is the predicted logits from the ego-classifier M^{ego}, and $i \in \{0, 1\}$ is the class label.

Third-person videos display various egocentric cues, resulting in a broad distribution of values for Ego-Score, despite sharing the same viewpoint (details in Supp). This score is used as the soft target in the auxiliary task loss, which we predict using a video classification head H^{ego}:

$$\mathbb{L}_{ego}(x) = -\sum_i y_i^{ego}(v)\ \log(H_i^{ego}(x)). \tag{2}$$

Object-Score: Finding interactive objects. In egocentric videos, interactions with objects are often central, as evident in popular egocentric video datasets [62, 12, 38]. Motivated by this, we designate an *Object-Score* task τ^{obj} for each video that encourages video representations to be predictive of manipulated objects.

Rather than require ground-truth object labels for third-person videos, we propose a simple solution that directly uses an off-the-shelf object recognition model M^{obj} trained on ImageNet [14] to describe objects in the video. Formally, for a video v with frames $\{f_1, ..., f_T\}$ we average the predicted logits from M^{obj} across frames to generate the video-level Object-Score $y_i^{obj}(v)$:

$$y_i^{obj}(v) = \frac{\exp(\frac{1}{T\beta}\sum_t z_i^{obj}(f_t))}{\sum_j \exp(\frac{1}{T\beta}\sum_t z_j^{obj}(f_t))}, \tag{3}$$

where $z_i^{obj}(f_t)$ is predicted logits for the i^{th} class from the recognition model, and β is the temperature parameter.

Similar to the Ego-Score, we introduce a knowledge-distillation loss during pre-training to make the video model predictive of the Object-Score using a module H^{obj}:

$$\mathbb{L}_{obj}(x) = -\sum_i y_i^{obj}(v) \log H_i^{obj}(x). \tag{4}$$

Interaction-Map: Discovering hand interaction regions. The Object-Score attempts to describe the interactive objects. Here we explicitly focus on the spatiotemporal regions of interactions in videos. Prior work shows it is possible to recognize a camera wearer's actions by attending to only a small region around the gaze point [38], as gaze often focuses on hand-object manipulation. Motivated by this, we introduce an *Interaction-Map* task τ^{int} to learn features that are predictive of these important spatiotemporal hand-object interaction regions in videos.

We adopt an off-the-shelf hand-object detector [60] M^{int} to detect hands and interacting objects. For each frame f_t in a video, the hand detector predicts a set of bounding-box coordinates and associated confidence scores $\mathcal{B}^t = \{(b^h, s^h)\}$ for detected hands. These bounding boxes are scaled to $h \times w$—the spatial dimensions of the video clip feature. We then generate a $t \times h \times w$ spatiotemporal

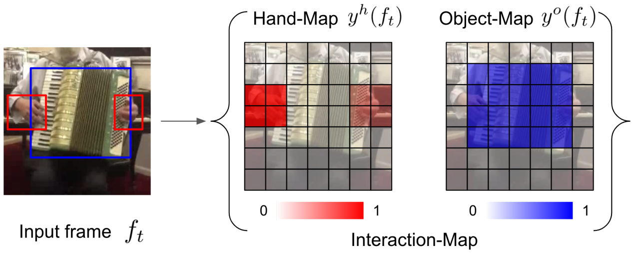

Figure 3: **Interaction-Map construction.** Soft-labels for hand-map and object-map are generated from detected bounding boxes according to Eqn 5.

hand-map y^h where the score for each grid-cell $y^h_{i,j}(f_t)$ at each time-step is calculated by its overlap with the detected hands at time t:

$$y^h_{i,j}(f_t) = \begin{cases} \max_{k \in \mathcal{B}^t_{i,j}} \{s^h_k\} & \text{if} \quad \mathcal{B}^t_{i,j} \neq \emptyset, \\ 0 & \text{if} \quad \mathcal{B}^t_{i,j} = \emptyset, \end{cases} \tag{5}$$

where $\mathcal{B}^t_{i,j}$ is the set of predicted bounding boxes that overlap with the (i, j)-th grid-cell at time t. The full hand-map label y^h is formed by concatenating the per-frame labels over time. We generate the corresponding object-map y^o analogously. See Fig 3 for an illustrative example.

We use the hand-map and object-map as the Interaction-Map pseudo-labels for the third-person videos during pre-training. We introduce two Interaction-Map prediction heads, H^h and H^o, to directly predict the $t \times h \times w$ Interaction-Map labels from clip features using a 3D convolution head:

$$\begin{aligned} \mathbb{L}_{int}(x) = -\sum_{i,j} \Bigg(& \sum_k y^h(f_k) \log[H^h(x)]_k \\ & + y^o(f_k) \log[H^o(x)]_k \Bigg)_{i,j}, \end{aligned} \tag{6}$$

where f_k is k-th frame from training clip x, and $H^h(x)$ and $H^o(x)$ are the predicted hand-map and object-map scores.

We predict Interaction-Maps instead of directly predicting bounding boxes via standard detection networks [58] for two reasons. First, detection architectures are not directly compatible with standard video backbones—they typically utilize specialized backbones and work well only with high resolution inputs. Second, predicting scores on a feature map is more aligned with our ultimate goal to improve the feature representation for egocentric video tasks, rather than train a precise detection model.

3.4. Ego-Exo training and evaluation

The three proposed ego-specific auxiliary tasks are combined together during the pre-training procedure to construct the final training loss:

$$\begin{aligned} \mathbb{L}(x) = & \mathbb{L}_{act}(x) + w_{ego} * \mathbb{L}_{ego}(x) \\ & + w_{obj} * \mathbb{L}_{obj}(x) + w_{int} * \mathbb{L}_{int}(x), \end{aligned} \tag{7}$$

where $\mathbb{L}_{act}$ is the standard cross-entropy loss for third-person action recognition, and w_{ego}, w_{obj} and w_{int} are the corresponding loss weights for the three auxiliary tasks, selected via cross-validation on downstream tasks.

Note that third-person video instances without hand-object interactions or salient interactive objects still contribute to the auxiliary loss terms, and are not ignored. Our distillation models approximate the *responses* of the pre-trained egocentric models as soft-targets instead of hard labels, offering valuable information about perceived egocentric cues, whether positive or negative for the actual label.

Training with our auxiliary losses results in features that are more suitable for downstream egocentric tasks, but it does not modify the network architecture itself. Consequently, after pre-training, our model can be directly used as a drop-in replacement for traditional video encoders, and it can be applied to various egocentric video tasks.

4. Experiments

Datasets. Our experiments use the following datasets.

- **Kinetics-400** [34] is a popular third-person video dataset containing ∼300k videos and spanning 400 human action classes. We use this dataset to pre-train all our models.
- **Charades-Ego** [62] has ∼68k instances spanning 157 activity classes. Each instance is a pair of videos corresponding to the same activity, recorded in the first and third-person perspective. Our method does not require this pairing, and succeeds even if no pairs exist (Supp.).
- **EPIC-Kitchens** [12] is an egocentric video dataset with videos of non-scripted daily activities in kitchens. It contains 55 hours of videos consisting of 39k action segments, annotated for interactions spanning 352 objects and 125 verbs. **EPIC-Kitchens-100** [11] extends this to 100 hours and 90k action segments, and is currently the largest annotated egocentric video dataset.

Due to its large scale and diverse coverage of actions, Kinetics has widely been adopted as the standard dataset for pre-training both first- and third-person video models [27, 19, 35, 11]. EPIC-Kitchens and Charades-Ego are two large and challenging egocentric video datasets that are the subject of recent benchmarks and challenges.

Evaluation metrics. We pre-train all models on Kinetics, and fine-tune on Charades-Ego (first-person only) and EPIC-Kitchens for activity recognition. Following standard practice, we report mean average precision (mAP) for Charades-Ego and top-1 and top-5 accuracy for EPIC.

Implementation details. We build our Ego-Exo framework on top of PySlowFast [18] and use SlowFast [19] video models as backbones with 8 input frames and stride 8.

We use a Slow-only ResNet50 architecture for all ablation experiments, and a SlowFast ResNet50/101 architecture for final results.

For our distillation heads H^{ego} and H^{obj} (Sec 3.3), we use a spatiotemporal pooling layer, followed by a linear classifier. We implement our Interaction-Map heads H^h and H^o as two 3D conv layers with kernel sizes 1×3×3 and 1×1×1, and ReLU activation.

For our combined loss function (Eqn 7), we set the loss weights w_{ego}, w_{obj} and w_{int} to $0.1, 0.5, 1.0$ respectively through cross-validation on the EPIC-Kitchens validation set (cross-validation on Charades-Ego suggested similar weights). The temperature parameter β in Eqn 1 and Eqn 3 is set to 1. Training schedule and optimization details can be found in Supp.

4.1. Ego-Exo pre-training

We compare our pre-training strategy to these methods:

- **Scratch** does not benefit from any pre-training. It is randomly initialized and directly fine-tuned on the target egocentric dataset.
- **Third-only** is pre-trained for activity labels on Kinetics 400 [34]. This represents the status-quo pre-training strategy for current video models.
- **First-only** is pre-trained for verb/noun labels on EPIC-Kitchens-100 [11], the largest publicly available egocentric dataset.
- **Domain-adapt** introduces a domain adaptation loss derived from gradients of a classifier trained to distinguish between first- and third-person video instances [23]. This strategy has been used in recent work to learn domain invariant features for third-person vs. drone footage [10].
- **Joint-embed** uses paired first- and third-person video data from Charades-Ego to learn viewpoint-invariant video models via standard triplet embedding losses [61]. We first pre-train this model with Kinetics to ensure that the model benefits from large-scale pre-training.
- **Ego-Exo** is pre-trained on Kinetics-400, but additionally incorporates the three auxiliary egocentric tasks (Sec 3.3) together with the original action classification loss, to learn egocentric-specific features during pre-training.

For this experiment, all models share the same backbone architecture (Slow-only, ResNet-50) and only the pre-training strategy is varied to ensure fair comparisons. *Domain Adapt* uses additional unlabeled egocentric data during pre-training, but from the same target dataset that the model will have access to during fine-tuning. *Joint-embed* uses paired egocentric and third-person data, an advantage that the other methods do not have, but offers insight into performance in this setting. Only *First-only* has access to ego-videos labeled for actions during pre-training.

Methods	C-Ego mAP	EPIC verbs top-1	EPIC verbs top-5	EPIC nouns top-1	EPIC nouns top-5
Scratch	8.50	55.62	86.10	38.35	62.39
First-only	11.73	–	–	–	–
Third-only	24.69	61.19	87.49	46.18	69.72
Domain-adapt [23]	23.62	61.27	87.49	45.93	68.73
Joint-embed [61]	-	61.26	87.17	46.55	68.97
Ego-Exo	**26.23**	**62.83**	**87.63**	**48.15**	**70.28**

Table 1: **Ego-Exo vs. alternate pre-training methods.** Our Ego-Exo pre-training results in best performance. Note that we do not evaluate *First-only* and *Joint-embed* on the datasets they were pre-trained on (EPIC and Charades-Ego respectively). Values are averaged over 3 runs.

Methods	τ^{ego}	τ^{obj}	τ^{h+o}	C-Ego mAP	EPIC verbs top-1	EPIC verbs top-5	EPIC nouns top-1	EPIC nouns top-5
Third-only	✗	✗	✗	24.69	61.19	87.49	46.18	69.72
Ego-Exo	✓	✗	✗	25.01	62.22	87.78	46.26	68.76
	✗	✓	✗	25.49	61.65	87.57	46.27	69.52
	✗	✗	✓	25.91	62.55	**88.50**	47.71	69.62
	✓	✓	✓	**26.23**	**62.83**	87.63	**48.15**	**70.28**

Table 2: **Auxiliary task ablation study.** Distilling knowledge from all three egocentric tasks results in the best performing pre-trained model. Values are averaged over 3 runs.

Table 1 shows the validation performance of different pre-training strategies. *Third-only* benefits from strong initialization from large-scale video pre-training and greatly outperforms models trained from *Scratch*. *First-only* performs very poorly despite being pre-trained on the largest available egocentric dataset, indicating that increasing scale alone is not sufficient—the diversity of scenes and activities in third-person data plays a significant role in feature learning as well. *Domain-adapt* and *Joint-embed* both learn viewpoint invariant features using additional unlabeled egocentric data. However, the large domain gap and small scale of the paired dataset limit improvements over *Third-only*. Our Ego-Exo method achieves the best results on both Charades-Ego and EPIC-Kitchens. The consistent improvements (+1.54% mAP on Charades-Ego, and +1.64%/+1.97% on EPIC verbs/nouns) over Third-only demonstrate the effectiveness of our proposed auxiliary egocentric tasks during pre-training. This is a key result showing the impact of our idea.

Fig 4 shows a class-wise breakdown of performance on Charades-Ego compared to the *Third-only* baseline. Our pre-training strategy results in larger improvements on classes that involve active object manipulations.

4.2. Ablation studies

Impact of auxiliary ego-tasks. We next analyze the impact of each auxiliary egocentric task in our Ego-Exo framework. As shown in Table 2, adding the *Ego-Score* task improves performance on both EPIC-Kitchens tasks, while

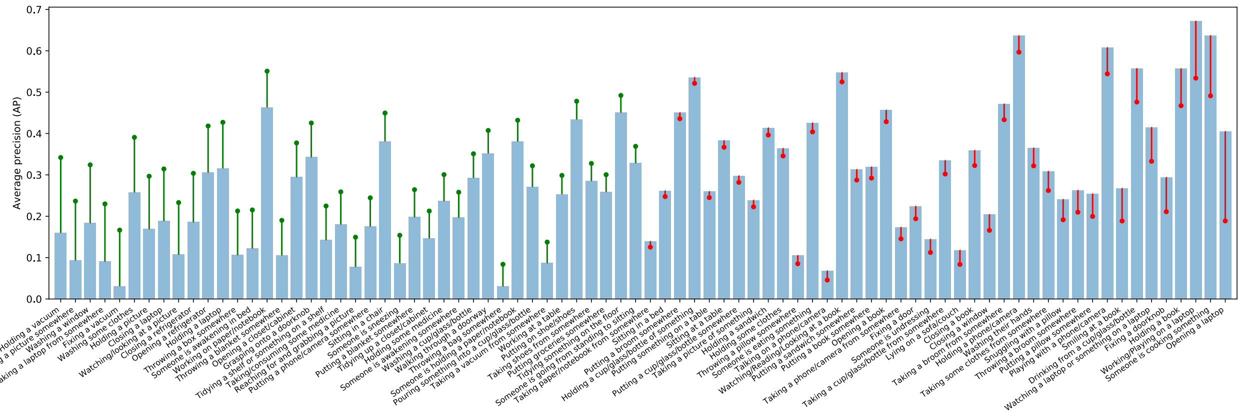

Figure 4: **Class-level performance on Charades-Ego.** Our method significantly improves on several classes that focus on active object manipulations (green lines), and performs only marginally worse across most under-performing classes (red lines). 35 most/least improved classes (out of all 157 classes) are shown.

	C-Ego	EPIC verbs		EPIC nouns	
Methods	mAP	top-1	top-5	top-1	top-5
Third-only	24.69	61.19	87.49	46.18	69.72
Third-only +aux	25.00	62.36	87.72	46.59	68.33
Δ	*+0.31*	*+1.17*	*+0.23*	***+0.42***	*-1.39*
Ego-Exo	26.23	62.83	87.63	48.15	70.28
Ego-Exo +aux	27.47	64.26	88.45	48.39	70.68
Δ	***+1.24***	***+1.43***	***+0.82***	*+0.24*	***+0.40***

Table 3: **Effect of Ego-Exo losses during fine-tuning.** Adding distillation losses during fine-tuning improves performance for both models, and results in a larger performance gain for our Ego-Exo pre-trained models. Values are averaged over 3 runs.

Method	mAP
ActorObserverNet [61]	20.0
SSDA [10]	23.1
I3D [10]	25.8
SlowFast [19]	25.93
Ego-Exo	28.04
Ego-Exo*	29.19
Ego-Exo*-R101	**30.13**

Table 4: **Comparison to prior work on Charades-Ego.** Despite having no access to paired egocentric data, our model outperforms specialized joint-embedding and domain adaptation based methods.

adding *Object-Score* and *Interaction-Map* consistently improves all results. This reveals that despite varying structure and semantics, these scores capture important underlying egocentric information to complement third-person pre-training, and further boost performance when used together.

Fig 5 shows instances from Kinetics based on our auxiliary pseudo-label scores combined with the weights in Eqn 7. Our score is highest for object-interaction heavy activities (e.g., top row: knitting, changing a tire), while it is low for videos of broader scene-level activities (e.g., bottom row: sporting events). Note that these videos are not in the egocentric viewpoint—they are largely third-person videos from static cameras, but are *ego-like* in that they prominently highlight important features of egocentric activity (e.g. hands, object interactions).

Adding auxiliary ego-tasks during fine-tuning. Our auxiliary losses may also be added *after* pre-training, for fine-tuning downstream egocentric models similar to prior semi-supervised learning work [9]. We re-introduce our Interaction-Map loss $\mathbb{L}_{int}$ (Eqn 6) for downstream egocentric training. We do not include Ego-Score (which is trivially high for all videos) and Object-Score (which is subsumed in the interaction label for this setting) as their impact after pre-training was minimal.

Table 3 shows that while both the baseline and our method further improve by adding the auxiliary task during fine-tuning, our improvements (Ego-Exo + aux) are larger, especially on Charades-Ego. This is likely because our distillation heads benefit from training to detect hands and objects in large-scale third-person video prior to fine-tuning for the same task on downstream ego-datasets.

4.3. Comparison with state-of-the-art

Finally, we compare our method with state-of-the-art models, many of which use additional modalities (flow, audio) compared to our RGB-only models. We include three competitive variants of our model using SlowFast [19] backbones: (1) **Ego-Exo** uses a ResNet50 backbone; (2) **Ego-Exo*** additionally incorporates our auxiliary distillation loss during fine-tuning.[1] (3) **Ego-Exo*-R101** further uses a ResNet-101 backbone.

[1] Same as Ego-Exo + aux in Table 3, but here with a SlowFast backbone

	Methods	Overall						Unseen Participants			Tail Classes		
		top-1			top-5			top-1			top-1		
		verb	noun	actions	verb	noun	actions	verb	noun	actions	verb	noun	actions
w/ audio	Epic-Fusion [35]	62.40	46.50	35.11	88.74	72.24	55.17	56.57	41.78	29.46	29.41	18.55	13.47
w/o audio	TSN fusion [12]	58.43	46.54	32.79	87.27	72.49	53.12	52.04	42.09	26.30	24.76	14.91	10.41
	TRN [22]	62.56	45.70	34.41	88.24	71.37	54.65	57.49	40.85	28.69	27.24	13.42	11.20
	TSM [40]	65.51	48.48	37.58	**89.39**	73.46	58.04	59.66	43.16	30.41	29.76	15.84	13.15
	SlowFast [19]	63.89	49.66	37.42	88.71	74.99	58.17	57.37	44.31	29.71	33.57	22.57	16.55
	Ego-Exo	64.08	50.59	38.38	88.57	74.82	59.04	57.42	46.18	31.41	32.87	**22.98**	16.30
	Ego-Exo*	65.02	**51.74**	39.52	89.26	75.95	60.07	58.86	**47.01**	32.36	33.68	22.35	16.30
	Ego-Exo*-R101	**66.07**	51.51	**39.98**	**89.39**	**76.31**	**60.68**	**59.83**	45.50	**32.63**	**33.92**	22.91	**16.96**

Table 5: **Comparison on EPIC-Kitchens-100 action recognition test set.** Our method is best in all categories.

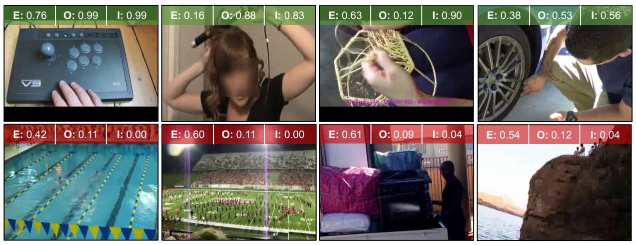

Figure 5: **Kinetics instances sorted by Ego-Exo scores.** Our task scores are maximum for videos that prominently feature hands/objects in view (top row), and minimum for scenes devoid of human-centered activity (bottom row).

Charades-Ego. Table 4 compares our Ego-Exo method with existing methods. Our Ego-Exo and Ego-Exo* yield state of the art accuracy, improving performance over the strongest baseline by +2.11% and +3.26% mAP. We observe large performance gains over prior work, including ActorObserverNet [61] and SSDA [10], which use joint-embedding or domain adaptation approaches to transfer third-person video features to the first-person domain. In addition, unlike the competing methods, our method does not require any egocentric data which is paired or shared category labels with third-person data during pre-training.

EPIC-Kitchens. Table 6 compares our method to state-of-the-art models on the EPIC-Kitchens test set. Ego-Exo and Ego-Exo* consistently improve over SlowFast (which shares the same backbone architecture) for all categories on both seen and unseen test sets. Epic-Fusion [35] uses additional optical flow and audio modalities together with RGB, yet Ego-Exo outperforms it on the top-1 metric for all categories. AVSlowFast [71] also utilizes audio, but is outperformed by our model with the same backbone (Ego-Exo*-R101) on the S1 test set.[2] On **EPIC-Kitchen-100** [11], as shown in Table 5, Ego-Exo consistently improves over the SlowFast baseline on all evaluation metrics, and Ego-Exo*-R101 outperforms *all* existing state-of-the-art methods.

[2]Table 6 compares existing methods under a controlled setting: using a single model with RGB or RGB+audio as input, and only Kinetics/ImageNet for pre-training. Other reported results on the competition page may use extra modalities, larger pre-training datasets, or model ensemble schemes (e.g. the top ranking method ensembles 8 models).

S1 (seen)	Methods	verbs top-1	verbs top-5	nouns top-1	nouns top-5	actions top-1	actions top-5
w/ audio	Epic-Fusion [35]	64.75	**90.70**	46.03	71.34	34.80	56.65
	AVSF-R101 [71]	**65.70**	89.50	**46.40**	**71.70**	**35.90**	**57.80**
w/o audio	TSN fusion [12]	48.23	84.09	36.71	62.32	20.54	39.79
	RU-LSTM [22]	56.93	85.68	43.05	67.12	33.06	55.32
	SlowFast [19]	64.57	89.67	45.89	69.50	34.67	54.47
	Ego-Exo	65.84	89.91	46.68	70.30	36.00	54.90
	Ego-Exo*	**66.19**	90.11	47.21	70.08	37.04	55.91
	Ego-Exo*-R101	65.97	**90.32**	**47.99**	**70.72**	**37.09**	**56.32**
S2 (unseen)							
w/ audio	Epic-Fusion [35]	52.69	79.93	27.86	53.87	19.06	36.54
	AVSF-R101 [71]	**55.80**	**81.70**	**32.70**	**58.90**	**24.00**	**43.20**
w/o audio	TSN fusion [12]	39.40	74.29	22.70	45.72	10.89	25.26
	RU-LSTM [22]	43.67	73.30	26.77	48.28	19.49	37.15
	SlowFast [19]	53.91	80.81	30.15	55.48	21.58	37.56
	Ego-Exo	54.11	80.37	30.83	57.08	21.20	38.34
	Ego-Exo*	53.88	80.51	31.44	57.60	22.02	39.13
	Ego-Exo*-R101	**55.34**	**81.46**	**31.72**	**58.25**	**22.81**	**40.18**

Table 6: **Comparison to prior work on EPIC-Kitchens (test set).** Methods in gray use additional audio modality information. Our method outperforms all methods that use consistent modalities in both settings, and is competitive with models that benefit from audio stream inputs.

5. Conclusion

We proposed a novel method to embed key egocentric signals into the traditional third-person video pre-training pipeline, so that models could benefit from both the scale and diversity of third-person video datasets, and create strong video representations for downstream egocentric understanding tasks. Our experiments show the viability of our approach as a drop-in replacement for the standard Kinetics-pretrained video model, achieving state-of-the-art results on egocentric action recognition on Charades-Ego and EPIC-Kitchens-100. Future work could explore alternate distillation tasks and instance-specific distillation losses to maximize the impact of third-person data for training egocentric video models.

Acknowledgments: UT Austin is supported in part by the NSF AI Institute and FB Cognitive Science Consortium.

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *CVPR*, 2018. 1, 2

[2] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *ECCV*, 2016. 2

[3] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171, 2018. 2

[4] Shervin Ardeshir and Ali Borji. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *ECCV*, 2018. 2

[5] Shervin Ardeshir, Sandesh Sharma, and Ali Broji. Egoreid: Cross-view self-identification and human re-identification in egocentric and surveillance videos. *arXiv preprint arXiv:1612.08153*, 2016. 2

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, 2016. 2

[7] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *ECCV*, 2018. 2

[8] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NeurIPS*, 2017. 2

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 7

[10] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *WACV*, 2020. 1, 2, 6, 7, 8

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020. 1, 2, 5, 6, 8, 12, 14

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 2, 4, 5, 8, 12

[13] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1), 2016. 1

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[15] Eadom Dessalene, Michael Maynord, Chinmaya Devaraj, Cornelia Fermuller, and Yiannis Aloimonos. Egocentric object manipulation graphs. *arXiv preprint arXiv:2006.03201*, 2020. 2

[16] Mohamed Elfeki, Krishna Regmi, Shervin Ardeshir, and Ali Borji. From third person to first person: Dataset and baselines for synthesis and retrieval. *arXiv preprint arXiv:1812.00104*, 2018. 2

[17] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *CVPR*, 2017. 2

[18] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 5

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 1, 5, 7, 8, 12, 13

[20] A. Furnari, S. Battiato, K. Grauman, and G. Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 2017. 2

[21] Antonino Furnari and Giovanni Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2

[22] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *ICCV*, 2019. 2, 8

[23] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1), 2016. 6

[24] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020. 2

[25] Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. Distinit: Learning video representations without a single labeled video. In *ICCV*, 2019. 2

[26] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "Something Something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 13

[27] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 5

[28] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, 2016. 2

[29] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[30] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *ECCV*, 2018. 2

[31] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *ECCV*, 2018. 2

[32] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *CVPR*, 2017. 2
[33] Georgios Kapidis, Ronald Poppe, Elsbeth Van Dam, Lucas Noldus, and Remco Veltkamp. Egocentric hand track and object-based human action recognition. In *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2019. 2
[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 5, 6
[35] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019. 1, 2, 5, 8, 13
[36] Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summarization. *IJCV*, 2015. 1, 2
[37] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 2
[38] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, 2018. 2, 4
[39] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *CVPR*, 2015. 2
[40] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 8
[41] Gaowen Liu, Hao Tang, Hugo Latapie, and Yan Yan. Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP*, 2020. 2
[42] Miao Liu, Siyu Tang, Yin Li, and James Rehg. Forecasting human object interaction: Joint prediction of motor attention and egocentric activity. *arXiv preprint arXiv:1911.10967*, 2019. 2
[43] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015. 2
[44] Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *ICCV Workshops*, 2019. 2
[45] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *ECCV*, 2018. 2
[46] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016. 2
[47] Stefan Mathe and Cristian Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *ECCV*, 2012. 2
[48] Andrew N Meltzoff. Imitation of televised models by infants. *Child development*, 59(5), 1988. 1
[49] Ravi Teja Mullapudi, Steven Chen, Keyi Zhang, Deva Ramanan, and Kayvon Fatahalian. Online model distillation for efficient video inference. In *ICCV*, 2019. 2
[50] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *CVPR*, 2020. 2
[51] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*, 2020. 2
[52] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, 2020. 2
[53] Fiora Pirri, Lorenzo Mauro, Edoardo Alati, Valsamis Ntouskos, Mahdieh Izadpanahkakhk, and Elham Omrani. Anticipation and next action forecasting in video: an end-to-end model with memory. *arXiv preprint arXiv:1901.03728*, 2019. 2
[54] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 2
[55] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4), 1978. 1
[56] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *CVPR*, 2018. 2
[57] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *ICCV*, 2019. 2
[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 5
[59] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *ICCV*, 2017. 2
[60] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 2, 4, 12
[61] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, 2018. 1, 2, 6, 7, 8
[62] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 1, 2, 4, 5, 12, 15
[63] Bilge Soran, Ali Farhadi, and Linda Shapiro. Action recognition in the presence of one egocentric and multiple static cameras. In *ACCV*, 2014. 2
[64] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. In *WACV*, 2020. 2
[65] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *CVPR*, 2019. 2
[66] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019. 2
[67] 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 *ECCV*, 2016. 1

[68] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020. 2

[69] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 1

[70] Xiaohan Wang, Linchao Zhu, Yu Wu, and Yi Yang. Symbiotic attention for egocentric action recognition with object-centric alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[71] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 8

[72] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In *ECCV*, 2018. 2

[73] Liang Yang, Hao Jiang, Jizhong Xiao, and Zhouyuan Huo. Ego-downward and ambient video based person location association. *arXiv preprint arXiv:1812.00477*, 2018. 2

[74] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Visual motif discovery via first-person vision. In *ECCV*, 2016. 1, 2

[75] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *ACM MM*, 2019. 1, 2

[76] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 1

[77] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *ICCV*, 2015. 1, 2