Ego-Only: Egocentric Action Detection without Exocentric Transferring

Huiyu Wang[1] Mitesh Kumar Singh[1] Lorenzo Torresani[1]
[1]Meta AI

Abstract

*We present Ego-Only, the **first** approach that enables state-of-the-art action detection on egocentric (first-person) videos without any form of exocentric (third-person) transferring. Despite the content and appearance gap separating the two domains, large-scale exocentric transferring has been the default choice for egocentric action detection. This is because prior works found that egocentric models are difficult to train from scratch and that transferring from exocentric representations leads to improved accuracy. However, in this paper, we revisit this common belief. Motivated by the large gap separating the two domains, we propose a strategy that enables effective training of egocentric models without exocentric transferring. Our Ego-Only approach is simple. It trains the video representation with a masked autoencoder finetuned for temporal segmentation. The learned features are then fed to an off-the-shelf temporal action localization method to detect actions. We find that this renders exocentric transferring unnecessary by showing remarkably strong results achieved by this simple Ego-Only approach on three established egocentric video datasets: Ego4D, EPIC-Kitchens-100, and Charades-Ego. On both action detection and action recognition, Ego-Only outperforms previous best exocentric transferring methods that use orders of magnitude more labels. Ego-Only sets new state-of-the-art results on these datasets and benchmarks without exocentric data.*

1. Introduction

In this paper we consider the problem of action detection from egocentric videos [30, 21, 19] captured by head-mounted devices. While action detection in third-person videos [6, 36] has been the topic of extended and active research by the computer vision community, the formulation of this task in the first-person setting is underexplored.

One major challenge of egocentric action detection is the lack of data, *i.e.* insufficient amount of egocentric videos to train large-capacity models to competitive results. For example, existing methods such as Ego-Exo [43] and Charades-Ego [56], attempted to train egocentric models from scratch using egocentric data only, but failed to obtain satisfactory results. Therefore, current egocentric action detection methods rely on out-of-domain large-scale exocentric (third-person) videos [37] or even images [22], under the assumption that the large-scale pretraining with proper transferring techniques can mitigate the negative effect of the domain gap between egocentric and exocentric videos. This hope is reinforced by the observation that deep neural networks exhibit invariance to object viewpoints [55],

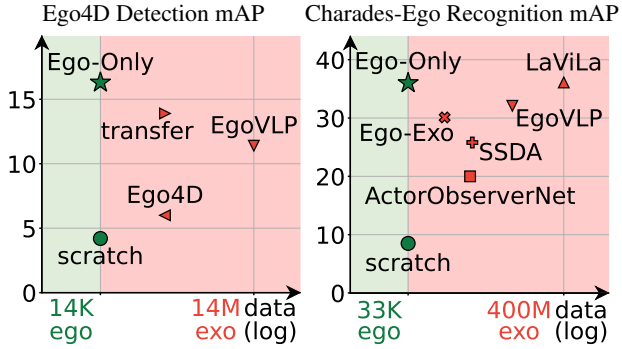

Figure 1. Our Ego-Only approach achieves state-of-the-art results on Ego4D [30] action detection and Charades-Ego [56] action recognition without any extra data or labels (Section 4). Compared with exocentric transferring, Ego-Only uses orders of magnitude fewer labels, simplifies the pipeline, and improves the results.

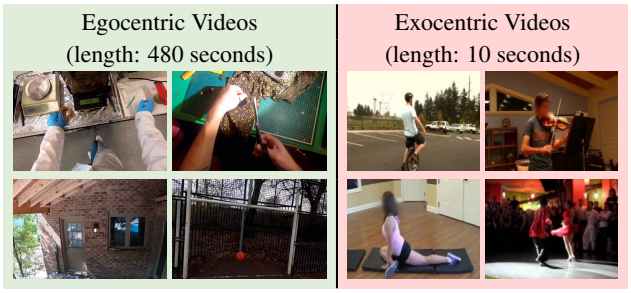

Figure 2. Domain gap between egocentric videos (Ego4D [30]) and exocentric videos (Kinetics-400 [37]). Exocentric videos are typically in the form of short trimmed clips, which show the actors as well as the contextual scene. Egocentric videos are dramatically longer, capture close-up object interactions but only the hands of the actor. These differences make it *challenging to transfer* models from exocentric action recognition to egocentric action detection.

as evidenced by the effective transfers from large-scale ImageNet pretraining to various still-image [50, 35, 62] and video understanding tasks [37, 5, 2]. Prior video approaches [43, 56] also demonstrated empirical benefits of transferring from exocentric representations over simply learning egocentric representations from scratch. As a result, this line of research focuses mainly on improving the transferring techniques that minimize the domain gap, or simply scaling exocentric data to a huge amount [72, 29].

However, we argue that the dramatically different viewpoint of first-person videos poses challenges that may not be addressed simply by scaling exocentric data or designing better transferring techniques, as illustrated in Figure 2: (1) No actor in view. In egocentric videos, the subject is behind the camera and is never visible, except for their hands. Conversely, third-person videos usually capture the actors as well as informative spatial context around them. (2) Domain shift. Egocentric videos entail daily life activities such as cooking, playing, performing household chores, which are poorly represented in third-person datasets. (3) Class granularity. First-person vision requires fine-grained recognition of actions within the same daily life category, such as "wipe oil metallic item", "wipe kitchen counter", "wipe kitchen appliance", and "wipe other surface or object" [30]. (4) Object interaction. Egocentric videos capture a lot of human-object interactions as a result of the first-person viewpoint. The scales and views of the objects are dramatically different than in exocentric videos. (5) Long-form. Egocentric videos are typically much longer than exocentric videos and thus require long-term reasoning of the human-object interactions rather than single frame classification. (6) Long-tail. Real-world long-tail distribution is often observed in egocentric datasets, as they are uncurated and thus reflect the in-the-wild true distribution of activities, which is far from uniform. (7) Localization. Egocentric action detection requires temporally sensitive representations which are difficult to obtain from third-person video classification on short and trimmed clips.

We argue that these challenges impede effective transfer from the exocentric to the egocentric domain and may actually cause detrimental biases when adapting third-person models to the first-person setting (as shown in Section 4). Therefore, instead of following the common transferring assumption, we revisit the old good idea of training with in-domain egocentric data only, but this time in light of the development of recent data-efficient training methods, such as masked autoencoders [32, 59, 27] as well as the scale growth of egocentric data collections (*e.g.*, the recently introduced Ego4D dataset [30]).

In this paper, we study the possibility of training with only egocentric video data by proposing a simple "Ego-Only" training approach. Specifically, Ego-Only consists of three training stages: (1) a masked autoencoder stage that bootstraps the backbone representation, (2) a simple finetuning stage that performs temporal semantic segmentation of egocentric actions, and (3) a final detection stage using an off-the-shelf temporal action detector, such as ActionFormer [73], without any modification. This approach enables us to train an egocentric action detector from random initialization without any exocentric videos or images.

Empirically, we evaluate Ego-Only on the three largest egocentric datasets, Ego4D [30], EPIC-Kitchens-100 [21], Charades-Ego [56], and two tasks, action detection and action recognition. Surprisingly, Ego-Only outperforms all previous results based on exocentric transferring, setting new state-of-the-art results, obtained for the first time without additional data. Specifically, Ego-Only advances the state-of-the-art results on Ego4D Moments Queries detection (+6.5% average mAP), EPIC-Kitchens-100 Action Detection (+5.5% on verbs and +6.2% on nouns), Charades-Ego action recognition (+3.1% mAP), and EPIC-Kitchens-100 action recognition (+1.1% top-1 accuracy on verbs).

In addition to the state-of-the-art comparison, we also noticed a few critical factors (as shown in Section 4) for the effectiveness of an Ego-Only approach: (1) dramatic performance deterioration when skipping either MAE pretraining or temporal segmentation finetuning; (2) importance of MAE pretraining on egocentric (as opposed to exocentric) data to learn the in-domain distribution; (3) criticality of long-term modeling for good accuracy; (4) the sensitivity to amount of unsupervised data; (5) surprising lack of performance gains by joint ego-exo pretraining or finetuning.

In summary, our contributions are four-fold:

- We propose the first Ego-Only method that trains egocentric action representations effectively without any form of exocentric data or transferring.
- We demonstrate that exocentric transferring is *not necessary* for state-of-the-art egocentric action detection.
- Ego-Only advances state-of-the-art results on both action detection and action recognition, evaluated on three large-scale egocentric datasets.
- Our empirical evaluation reveals several critical factors for the effectiveness of an Ego-Only approach.

2. Related Work

Action recognition methods learn to classify actions in trimmed video clips. Recent action recognition models include convolutional neural networks [60, 10, 64, 61, 65, 45, 28, 26] and vision transformers [24, 5, 25, 44, 2, 52]. The learned action representations are often used as features for downstream tasks.

Temporal action localization aims to detect action instances from long videos. Most methods [48, 47, 71, 75] detect actions using frozen video features from action recognition models. Recently, ActionFormer [73] models long-

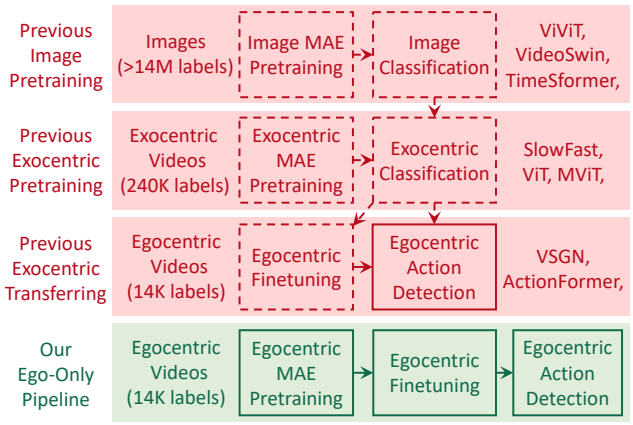

Figure 3. Our Ego-Only approach simplifies the previous pipeline by removing the dependence on pretrained exocentric checkpoints obtained with extra data, extra labels, and extra pretraining stages.

sequence features with transformers. SegTAD [74] detects actions via temporal segmentation. TALLFormer [17] trains the feature backbone end-to-end with the detector.

Self-supervised learning aims to learn visual representation without human annotation. Traditional methods include hand-crafted pretext tasks [54, 68, 39, 23] and contrastive learning [70, 33, 15, 31, 16, 7, 8, 9, 63]. Recently, masked autoencoders [4, 77, 32, 67, 27] have shown training efficiency [32], model scalability [32], data efficiency [59], and effectiveness on videos [67, 59, 27].

Egocentric video datasets [30, 20, 21, 57] have grown in size by orders of magnitude over the past few years, presenting new challenges [21] and opportunities[30], such as egocentric action recognition [21, 43] and detection [21]. Most egocentric action detection methods [30, 21, 73, 46] follow temporal action localization practices [75, 73, 47, 71] and adopt exocentric pretrained checkpoints [10, 28, 5, 3, 2].

In this paper, we study the possibility of detecting egocentric actions without any form of exocentric transferring.

3. Method

In Section 3.1, we provide an overview of our Ego-Only approach which enables egocentric action detection without relying on exocentric transferring. The proposed Ego-Only method consists of three training stages: a standard masked autoencoder (MAE) pretraining stage, an egocentric finetuning stage, which we present in Section 3.2, and finally standard training of a temporal action detector.

3.1. Ego-Only

There is an extensive literature about training object detectors [50, 35] on images end-to-end from random initialization [34]. However, these approaches are difficult to adapt to egocentric action detection where both the videos and the actions are long-form. For example, Ego4D [30] Moments clips are 8 minutes long, and around half of the actions are longer than 10 seconds which is the typical length of an exocentric video. In this case, end-to-end training of an action detector is impossible due to GPU memory limitations unless one reduces aggressively the model size, the spatial resolution, or the temporal sampling density, which would lead to degradation in performance.

This empirical challenge calls for a "proxy" objective that enables learning visual representations with a large model size, a high spatial resolution, and a high temporal sampling density. This surrogate objective is usually realized by pretraining on short exocentric videos. However, as discussed in Section 1, the learned representation may not transfer effectively. Instead, in our Ego-Only approach, we approximate the temporal action detection task by performing temporal semantic segmentation that predicts action labels at each frame. Note that this approximation is not exact because we truncate long-form videos into clips, throwing away the action context outside the sampled clip. Such approximation leads to a trade-off between the action context and the temporal sampling density, ablated in Section 4.3.

This simple surrogate objective allows us to train visual representations from random initialization towards temporal action detection. However, we empirically find that the learned representation generalizes poorly even with strong augmentation and regularization. In order to further improve generalization, we introduce an additional MAE pretraining stage which has been shown to yield strong generalization in the low-data regime [59]. This additional pretraining improves generalization as shown in Table 5.

Putting these pieces together, Figure 3 summarizes our complete Ego-Only method that includes the initial MAE pretraining, the egocentric finetuning task as an approximation of action detection, and the final temporal action detector that incorporates full context of the whole long-form video. This approach differs from existing methods in the absence of an exocentric pretraining stage that requires large-scale annotated exocentric videos or images. For example, most prior approaches pretrain egocentric models on Kinetics-400 (K400) with 240K annotated videos, while our Ego-Only method uses merely 14K annotated action segments on Ego4D and achieves better results (Table 5).

Next, we describe in more detail the initial MAE pretraining stage and the final action detection stage that are both adopted from existing literature without any modification. Note that this paper aims to revisit the value of exocentric transferring and does so by proposing an ego-only meta algorithm that is intentionally kept as simple as possible.

Masked Autoencoder. Our method applies the original MAE [32] and video MAE [27] algorithms. Specifically, we consider the vanilla vision transformers [24, 27], ViT-B and ViT-L, as our architectures, due to the native support by MAE. We do not consider convolutional architec-

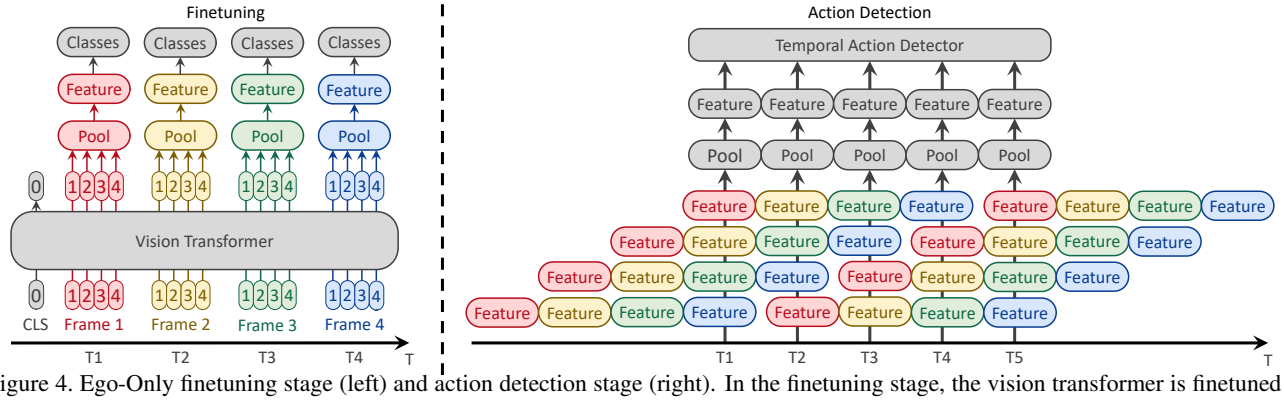

Figure 4. Ego-Only finetuning stage (left) and action detection stage (right). In the finetuning stage, the vision transformer is finetuned to predict action classes at each frame from spatially-pooled features (colors represent frame indices within a clip). In the detection stage, finetuned backbone features are frozen and extracted using a sliding window. Features at the same timestamp (*e.g.* T1) but from different windows are average-pooled. On top of the long sequence of frozen features, a detector is then trained to temporally localize the actions.

tures [28] or hierarchical transformers [51, 52, 25, 44] that require adaptation of the MAE algorithm. Since videos are highly redundant, we use a very high masking ratio (90%) with a random masking strategy and all of the pretraining recipes as suggested in video MAE [27]. The only adaptation we make is to sample each video with a probability proportional to its temporal length, because of the long-form property of egocentric videos. This ensures equal sampling probability for any possible clip in the dataset.

Action Detector. After the egocentric finetuning stage (Section 3.2) that trains the backbone representation towards action detection, we apply an existing temporal action localization algorithm to detect the actions. Specifically, given the finetuned video backbone, features are extracted from the frozen model with sliding windows, following standard practice in temporal action localization [73, 75]. Then, the action detector is trained on top of a long sequence of frozen video features to produce temporal segments as outputs. There is a potential risk of overfitting since our finetuning stage and action detection stage are trained on the same training set, but empirically we do not find this to be a significant issue in practice, probably because the detector takes as input a long-form video instead of a clip and the detector loss differs from simply segmentation. For better performance, we choose ActionFormer [73] as our default detector as it has demonstrated good accuracy on temporal action localization benchmarks. As we work on egocentric videos, we adopt the ActionFormer architecture previously proposed for EPIC-Kitchens-100 [21].

3.2. Finetuning via Temporal Segmentation

Inspired by TSN [64] and SegTAD [74] that detect actions via temporal semantic segmentation, we finetune our backbone features from MAE pretraining by predicting class labels for each frame, as illustrated in Figure 4 (left). This is akin to the task of image semantic segmentation [11, 12, 13, 14] which predicts class labels for each pixel. Formally, given an input video clip with a certain temporal span, a temporal segmentation model predicts output logits $L \in \mathbb{R}^{T \times C}$ where T denotes the temporal dimension of the logits and C is the total number of action classes.

We follow a few principles in defining this simple finetuning objective: (1) A video clip of a certain temporal span is taken as the input instead of the full long-form video. This temporal approximation enables us to train large-scale models within the given GPU memory limit. (2) We employ a fixed temporal span which is consistent with both MAE pretraining and detection feature extraction. This removes potential domain gaps when models are trained and inferred with different temporal spans. (3) The temporal segmentation objective trains models to distinguish frames of different classes within one video clip, especially when a long temporal span is adopted. (4) We train with clips uniformly sampled over the dataset, making full use of all positive and negative samples in the dataset.

Note that our segmentation stage differs from TSN [64] and SegTAD [74] mainly in the goal which is to finetune the backbone representation instead of to detect actions directly from the output scores. In order to address unique challenges (Section 1) in egocentric videos, we also adopt critical techniques addressing loss and imbalance issues.

Next, we discuss the loss function that we choose to finetune the backbone, how we address the egocentric imbalance challenges, and how backbone features are extracted for the subsequent action detection stage.

Loss function. Egocentric videos usually contain overlapping actions of different classes. For example, a person could be taking a photo while speaking on the phone. This makes the finetuning stage a multi-label classification task. Therefore, we employ a loss function independent for each action class, *i.e.* the activation of one class does not suppress another. Specifically, we adopt per-frame binary cross-entropy (BCE) as the loss function on the logits, instead of cross-entropy which suppresses non-maximum classes.

Imbalance challenges. The long-tail imbalance in egocentric videos (Section 1) poses a major challenge to our finetuning stage, due to the less curated nature and the long-form property of egocentric videos. Specifically, there are usually (1) imbalanced numbers of videos across action classes, (2) imbalanced action lengths within one class, and (3) imbalanced numbers of foreground frames vs background frames within one class. Inspired by the literature of one-stage object detection, we mitigate the imbalance issue by adopting focal loss [49] in the BCE loss and biasing the logits towards background at initialization. We also reweigh each action instance by the inverse of the action length, leading to a balanced loss for each instance.

Feature extraction. Once our video backbone is finetuned on sampled clips, features are extracted using a sliding window on both the training set and validation set for training the detector on long-form videos and validating the approach. According to temporal action localization literature [73, 75], clip features are average-pooled spatiotemporally following the exocentric classification practice [10, 28]. However, in our temporal segmentation case on long-form videos, our spatially-pooled features are trained to be temporally different within a video clip, encoding their own local context. Therefore, as illustrated in Figure 4 (right), given the sliding windows of features, we average-pool features at the same wall-clock timestamp from all sliding windows. This enables the usage of a long temporal span, such as 64 seconds (Figure 5), by extracting temporally variable features from a window.

4. Experiments

We evaluate our Ego-Only approach by reporting main results on the two largest egocentric video datasets, Ego4D [30] and EPIC-Kitchens-100 [21], measured by average mAP at tIoU $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ on the val set (Section 4.1). Next, we study the application to egocentric action recognition and report video-level mAP on Charades-Ego [56] and top-1 accuracy on EPIC-Kitchens-100 [21] (Section 4.2). Then, we carefully ablate the effect of each design choice in Section 4.3. Finally, we benchmark our model runtime cost in Section 4.4 and visualize egocentric MAE reconstructions in Section 4.5.

4.1. Main Results on Action Detection

Ego4D. We compare our results on the Ego4D [30] MQ val set with state-of-the-art methods in Table 1, using ViT-B and ViT-L. We notice that our Ego-Only performs significantly better than previous state-of-the-art but without any extra exocentric data or labels needed. Specifically, with ViT-B as the backbone, Ego-Only achieves an average mAP of 16.3%, producing a relative improvement of 170% over the Ego4D paper baseline [30] that pretrains on Kinetics-400 [37] with 18× annotated clips. This strong result even outperforms EgoVLP which has seen 4M language-narrated video clips from Ego4D (*i.e.* in-domain) and 14M images from IN-21K [22]. Finally, scaling Ego-Only to ViT-L backbone yields an mAP of 17.9%, setting a new state-of-the-art on this benchmark without any extra data or labels.

EPIC-Kitchens-100. Following the Ego4D exploration, we validate our Ego-Only approach on the EPIC-Kitchens-100 [21] Action Detection benchmark. We can see from Table 2 that Ego-Only achieves much better results compared with exocentric transferring. Specifically, compared with previous state-of-the-art methods that adopt Kinetics [37] SlowFast [28] features finetuned on EPIC-Kitchens-100 Action Recognition, our Ego-Only with a ViT-B backbone already performs 4.6% better on both verbs and nouns. Scaled to a ViT-L backbone, Ego-Only improves further and sets a new state-of-the-art result of 29.0% mAP on verbs and 28.1% mAP on nouns. By analyzing our results using DETAD [1] (supplementary material), we find that Ego-Only significantly reduces false positives on backgrounds, compared with exocentric transferring, probably because Kinetics contains mostly trimmed videos with foreground actions only. This validates the benefit of Ego-Only.

4.2. Application to Action Recognition

Besides egocentric action detection, we further evaluate our Ego-Only approach on the task of action recognition on Charades-Ego [56] and EPIC-Kitchens-100 [21]. This is simply achieved by skipping our last action detector stage and averaging the temporal semantic segmentation model output scores after the sigmoid activation in the BCE loss. Results on action recognition allow us to compare Ego-Only with a wider range of state-of-the-art methods.

Charades-Ego. In Table 3, we report recognition results on Charades-Ego [56] by finetuning the existing Ego4D MAE checkpoints on Charades-Ego, without exploiting any ego-exo supervision or correspondence. Remarkably, Ego-Only with a ViT-B backbone already significantly outperforms state-of-the-art methods that exploit ego-exo alignment (ActorObserverNet [56]), or semi-supervised domain adaptation (SSDA [18]), or ego-exo distillation (Ego-Exo [43]), or egocentric video-language pretraining (EgoVLP [46]). Furthermore, we compare LaViLa that uses CLIP initialization with 400M text-image pairs, 4M Ego4D narration-clip pairs, as well as the large language model GPT-2 XL. Our Ego-Only trained on only the egocentric subset of Charades-Ego, matches this result with merely 33K labels (around 0.01% of 404M) and a smaller ViT-L backbone. Finally, when we augment Ego-Only with the exocentric subset of Charades-Ego, we observe a significant gain of 3.1% absolute points over the LaViLa state-of-the-art.

method	backbone	params	extra data	extra labels	0.1	0.2	0.3	0.4	0.5	avg	# labels
Ego4D [30]	SlowFast [28]	63M	Kinetics-400 [37]	240K	9.10	7.16	5.76	4.62	3.41	6.03	254K
EgoVLP [46]	Frozen [3]	178M	IN-21K [22] + EgoClip [46]	18M	16.63	-	11.45	-	6.57	11.39	18M
Ego-Only	ViT-B	86M	-	-	22.5	19.3	16.0	13.1	10.6	16.3	**14K**
Ego-Only	ViT-L	304M	-	-	24.6	20.8	17.7	14.9	11.7	**17.9**	**14K**

Table 1. Ego4D action detection on MQ val set (see Table 5 and Table 6 for ablations).

method	backbone	extra data	extra labels	verb						noun						# labels seen
				0.1	0.2	0.3	0.4	0.5	avg	0.1	0.2	0.3	0.4	0.5	avg	
BMN [47, 21]	SlowFast	K400	240K	10.8	9.8	8.4	7.1	5.6	8.4	10.3	8.3	6.2	4.5	3.4	6.5	307K
G-TAD [71]	SlowFast	K400	240K	12.1	11.0	9.4	8.1	6.5	9.4	11.0	10.0	8.6	7.0	5.4	8.4	307K
ActionFormer	SlowFast	K400	240K	26.6	25.4	24.2	22.3	19.1	23.5	25.2	24.1	22.7	20.5	17.0	21.9	307K
Ego-Only	ViT-B	-	-	31.1	30.4	28.9	26.6	23.4	28.1	30.0	29.2	27.8	25.1	20.7	26.5	**67K**
Ego-Only	ViT-L	-	-	32.0	31.5	20.0	27.4	24.0	**29.0**	31.5	30.8	29.2	26.5	22.5	**28.1**	**67K**

Table 2. EPIC-Kitchens-100 Action Detection val set (see Table 8 for ablations).

method	backbone	params	mAP	# labels
ActorObserver [56]	ResNet-152	60M	20.0	1.4M
SSDA [18]	I3D	12M	25.8	1.6M
Ego-Exo [43]	SlowFast-R101	75M	30.1	0.3M
EgoVLP [46]	TSF-B	178M	32.1	18M
LaViLa [76]	TSF-B	178M	33.7	404M
Ego-Only	ViT-B	87M	33.3	**33K**
LaViLa [76]	TSF-L	528M	36.1	404M
Ego-Only	ViT-L	304M	36.0	**33K**
Ego-Only†	ViT-L	304M	**39.2**	**67K**

Table 3. Charades-Ego recognition. †with full Charades-Ego data.

method	variant	verb	noun
IPL [66]	I3D, K400	68.6	51.2
ViViT [2]	ViViT-L/16x2, IN-21k+K400	66.4	56.8
MoViNet [40]	MoViNet-A6, 120 frames	72.2	57.3
MTV [72]	MTV-B, WTS-60M, 280p	69.9	63.9
MTCN [38]	MFormer-HR, IN-21k+K400+VGG-Sound	70.7	62.1
Omnivore [29]	Swin-B, IN21k+IN-1k+K400+SUN	69.5	61.7
MeMViT [69]	MeMViT, 32×3, K600, 105.6 sec	71.4	60.3
LaViLa [76]	TSF-L, WebImageText+Ego4D	72.0	62.9
Ego-Only	ViT-L, 32 frames, 3.2 sec	**73.3**	59.4

Table 4. EPIC-Kitchens-100 action recognition top-1 accuracy.

EPIC-Kitchens-100. In Table 4, we report action recognition top-1 accuracies on EPIC-Kitchens-100 [21] by evaluating the EPIC-Kitchens-100 temporal segmentation model from Section 4.1. We compare Ego-Only with state-of-the-art methods exploiting large-scale image data (ViViT [2]), or web-scale text-image pairs (MTV [72], LaViLa [76]), or multimodal audio (MTCN [38]) depth (Omnivore [29]) supervision, or 32× temporal support (MeMViT [69]). In contrast, our Ego-Only using the 495 videos in EPIC-Kitchens-100 as the only source of supervision achieves the state-of-the-art results of 73.3% on verb classification, outperforming the existing best result by 1.1%. This validates the effectiveness of Ego-Only in capturing hand-object interactions from egocentric videos.

method	self-sup. MAE	sup. exo	sup. ego	Ego4D mAP	# labels seen
exo-sup	-	K400	Ego4D	13.9	254K (18×)
ours	Ego4D	-	Ego4D	**16.3**	**14K (1×)**
scratch	-	-	Ego4D	4.2	14K (1×)
exo-MAE	K400	-	Ego4D	13.4	14K (1×)
exo-FT	K400	K400	Ego4D	16.2	254K (18×)

Table 5. Varying the pretraining stage. Ego-Only outperforms exocentric transferring with much fewer labels (14K vs. 240K+14K).

4.3. Ablation Study

In order to analyze our Ego-Only approach, we compare Ego-Only with common exocentric transferring solutions and ablate the importance of each stage in Ego-Only. We also scale the amount of data consumed, the model sizes, as well as the number of pretraining epochs. We perform all ablation studies on egocentric action detection benchmarks.

Varying the pretraining stage. Table 5 reports our results with different pretraining stages. Compared with the common exocentric supervised baseline of 13.9% mAP, our Ego-Only with exactly the same backbone, the same fine-tuning, and the same detector, achieves the performance of 16.3% (+2.4%) mAP by using egocentric data only and with merely 14K labels, instead of 240K labels used in the exocentric transferring method.

method	backbone	self-sup. MAE	sup. exo	sup. ego	Ego4D mAP
exo-sup	SlowFast	-	K400	-	13.2
exo-MAE	ViT-B	K400	-	-	6.7
ego-MAE	ViT-B	Ego4D	-	-	7.8
exo-FT	ViT-B	K400	K400	-	13.5
ours	ViT-B	Ego4D	-	Ego4D	16.3

Table 6. Varying the finetuning stage.

Next, we consider skipping the MAE pretraining and train from scratch the model via temporal segmentation on Ego4D. However, our best model learned from scratch only reaches the mAP of 4.2% (vs. 16.3% with MAE pretraining in Ego-Only), due to the limited number of labels available on Ego4D, only 14K. This is smaller than the number of labels in MNIST [42] or CIFAR [41] but the task of egocentric action detection is significantly more challenging.

In addition to the model trained from scratch, we also compare with self-supervised MAE pretraining on Kinetics-400. When this checkpoint is finetuned, it achieves 13.4% mAP which is 2.9% worse than the counterpart pretrained on Ego4D. This gap is reasonable since the model is pretrained on out-of-domain data but does not benefit from the large-scale exocentric labels. Once the extra labels are used, Kinetics finetuning yields performance on-par with our much simpler Ego-Only approach.

Varying the finetuning stage. After varying the pretraining stage, we study the importance of finetuning. For this purpose, we extract features from pretrained models, without any form of finetuning on egocentric data. Contrary to the strong linear probing results of MAE on ImageNet-1K[22], we observe that frozen MAE features perform poorly on egocentric action detection, leading to an absolute drop of 8.5% points in average mAP. Kinetics-400 MAE features perform even worse (as expected), but finetuning on Kinetics with 240K labels is helpful, achieving a 13.5% mAP which is 2.8% worse than Ego-Only. We also try concatenating frozen MAE features from multiple blocks [9], but only observe a marginal gain (supplementary material).

Detectors and temporal spans. Next, we compare temporal action detector choices in Ego-Only and vary the temporal span at the same time. As we use a consistent temporal span for the whole pipeline, including MAE, finetuning, and feature extraction (Section 3.2), we pretrain MAE with each temporal span for 200 epochs only. Then, we define a simple baseline of a 1D blob detector [53] using the Laplacian of Gaussian kernel. To our surprise, as shown in Figure 5, this simple blob detection baseline achieves 8.2% mAP which is already better than the Ego4D [30] paper baseline of 6.0% mAP with pretrained SlowFast [28] features and VSGN [75], thanks to the effectiveness of Ego-Only features. We also notice that the blob detector and

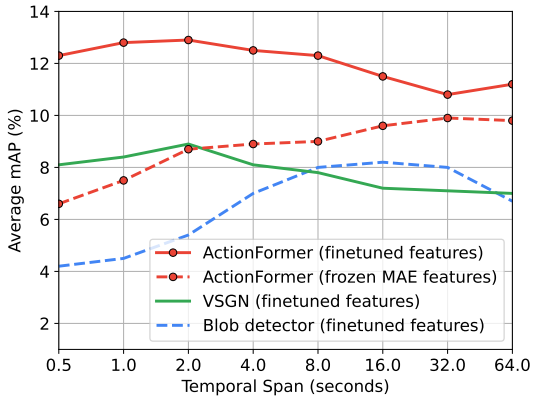

Figure 5. Varying detectors and temporal spans. The blob detector performs surprisingly well and perfers a long temporal span, while ActionFormer and VSGN prefer short spans due to their transformer or graph neural network based architectures.

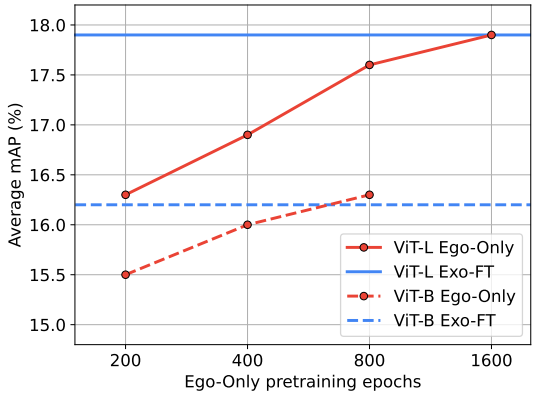

Figure 6. Scaling models and pretraining epochs. At around 800 or 1600 epochs, our Ego-Only starts to match exocentric transferring.

the frozen MAE feature prefer a longer temporal span of 16 or 32 seconds, demonstrating the importance of long-term context in egocentric videos. On the other hand, VSGN [75] and ActionFormer [73] prefer short feature spans probably because the graph neural network or the transformer captures long-term relations internally, benefiting more from local features that represent dense temporal motion. Finally, ActionFormer with finetuned features achieves the best result of 12.9%, outperforming VSGN by 4.0% consistently.

Scaling models and pretraining epochs. In addition to ablating the three stages in our Ego-Only pipeline, we also scale the model size from ViT-B to ViT-L and benchmark results under different computation budgets. We keep the relatively cheap finetuning of 20 epochs unchanged, but vary the MAE pretraining epochs. As shown in Figure 6, both ViT-B and ViT-L results improve consistently when they are pretrained longer. At around the budget of 800 or 1600 epochs, our Ego-Only models start to match Kinetics-400 pretrained models with both ViT-B and ViT-L. The Kinetics baselines, before transferred to egocentric data, are

ego MAE pretrain (hours)	ego finetune	mAP
random initialization (0h)	195h	4.2
Ego4D MQ clips (195h)	195h	14.5
Ego4D MQ videos (487h)	195h	14.8
Ego4D EM videos (838h)	195h	14.7
Ego4D ALL videos (3560h)	195h	15.5

Table 7. Scaling the amount of pretraining data. **MQ clips**: all MQ training clips [30]. **MQ videos**: all videos in the MQ task training set. **EM videos**: all videos in the Episodic Memory benchmark training set. **ALL videos**: all Ego4D videos except MQ val and test videos. Our Ego-Only results improve with respect to the amount of data consumed in the pretraining stage.

method	self-sup. MAE	sup. exo	sup. ego	verb mAP	noun mAP	# labels seen
frozen	K400	K400	-	17.9	14.6	307K
exo-FT	K400	K400	EPIC	28.0	28.3	307K
frozen	K600	K600	-	17.0	15.0	457K
exo-FT	K600	K600	EPIC	27.1	**28.6**	457K
ours	EPIC	-	EPIC	**29.0**	28.1	**67K**

Table 8. Scaling exocentric pretraining data.

pretrained with 800/1600 epoch MAE and 150/100 epoch exocentric finetuning that consumes not only more data and labels but also more computation resources than Ego-Only.

Scaling egocentric pretraining data. Beyond standard ablations on pretraining epochs, an intriguing dimension for study offered by the massive scale of Ego4D is the different amounts of large-scale unsupervised video data. Specifically, given the fixed amount of finetuning data, we select four subsets and amounts of unsupervised data in Ego4D to study the data scaling property of the Ego-Only pretraining stage. Note that in all cases, we exclude val and test videos of the MQ task from the pretraining set. All models are pretrained for 200 epochs instead of 800 epochs to save computation resources. From the results in Table 7, we see that the performance of Ego-Only improves as more unsupervised data is provided for MAE pretraining.

Scaling exocentric pretraining data. Besides scaling egocentric data, we study the common practice of scaling exocentric pretraining from K400 (240K videos) to K600 (390K videos). As shown in Table 8, scaling exocentric data improves noun mAP marginally and hurts verb mAP by 0.9%, compared with transferring from K400. This is probably due to the bias of Kinetics towards scene and object classification. When we evaluate on verbs, Ego-Only shows a significant absolute gain of 1.9% over K600 transferring that requires much more labels. This observation is also consistent with the action recognition results in Table 4, where Ego-Only achieves the state-of-the-art verb accuracy.

method	self-sup. MAE	sup. exo	sup. ego	Ego4D mAP
exo-MAE	K400	-	Ego4D	13.4
joint-MAE	K400 & Ego4D	-	Ego4D	16.0
ours	Ego4D	-	Ego4D	**16.3**

Table 9. Joint ego-exo pretraining.

method	self-sup. MAE	sup. exo	sup. ego	verb mAP	noun mAP	# labels seen
joint-FT	EPIC	-	KEEC	28.4	27.9	515K
ours	EPIC	-	EPIC	**29.0**	**28.1**	**67K**

Table 10. Joint ego-exo finetuning. **KEEC**: joint finetuning on **K**inetics-600, **E**go4D, **E**PIC-Kitchens-100, **C**OIN.

method	FLOPs (G)			training time (hours)			
	MAE	exo	ego	MAE	exo	ego	total
exo-sup	-	598	598	-	200.9	10.0	210.9
exo-FT	81	598	598	100.5	50.2	10.0	160.7
ours	81	-	598	100.5	-	10.0	**110.5**

Table 11. Inference FLOPs and training time for each stage. Our Ego-Only method reduces the total training cost by a large margin.

Joint ego-exo pretraining. In Table 9, we study the effect of joint ego-exo pretraining by building a joint-MAE variant that trains the MAE model on both K400 and Ego4D, instead of K400 or Ego4D individually. We observe that the results are greatly improved compared with the out-of-domain K400 transferring, but lags behind our Ego-Only.

Joint ego-exo finetuning. In Table 10, we explore joint ego-exo finetuning with a shared model backbone on four large-scale video datasets, including Kinetics-600, Ego4D, EPIC-Kitchens-100, and COIN [58]. This joint dataset contains 515K labeled clips, 7× more than our default finetuning data of 67K, but does not lead to any performance gain probably due to the domain gap between these datasets.

4.4. Runtime

Table 11 reports inference FLOPs and training time for each stage, on 64 V100s with ViT-L and 800-epoch MAE. Our MAE stage is identical to Video MAE [27]. We see that Ego-Only accelerates training significantly.

4.5. Visualization of MAE Reconstructions

In Figure 7, we visualize MAE [32, 27] reconstruction results on Ego4D [30] with a ViT-B [24] trained for 200 epochs without per-patch normalization. We notice that egocentric MAE learns human-object interactions (a,b,c,d) and temporal correspondence across frames (e,f), even in cases with strong head/camera motion (g,h,i,j).

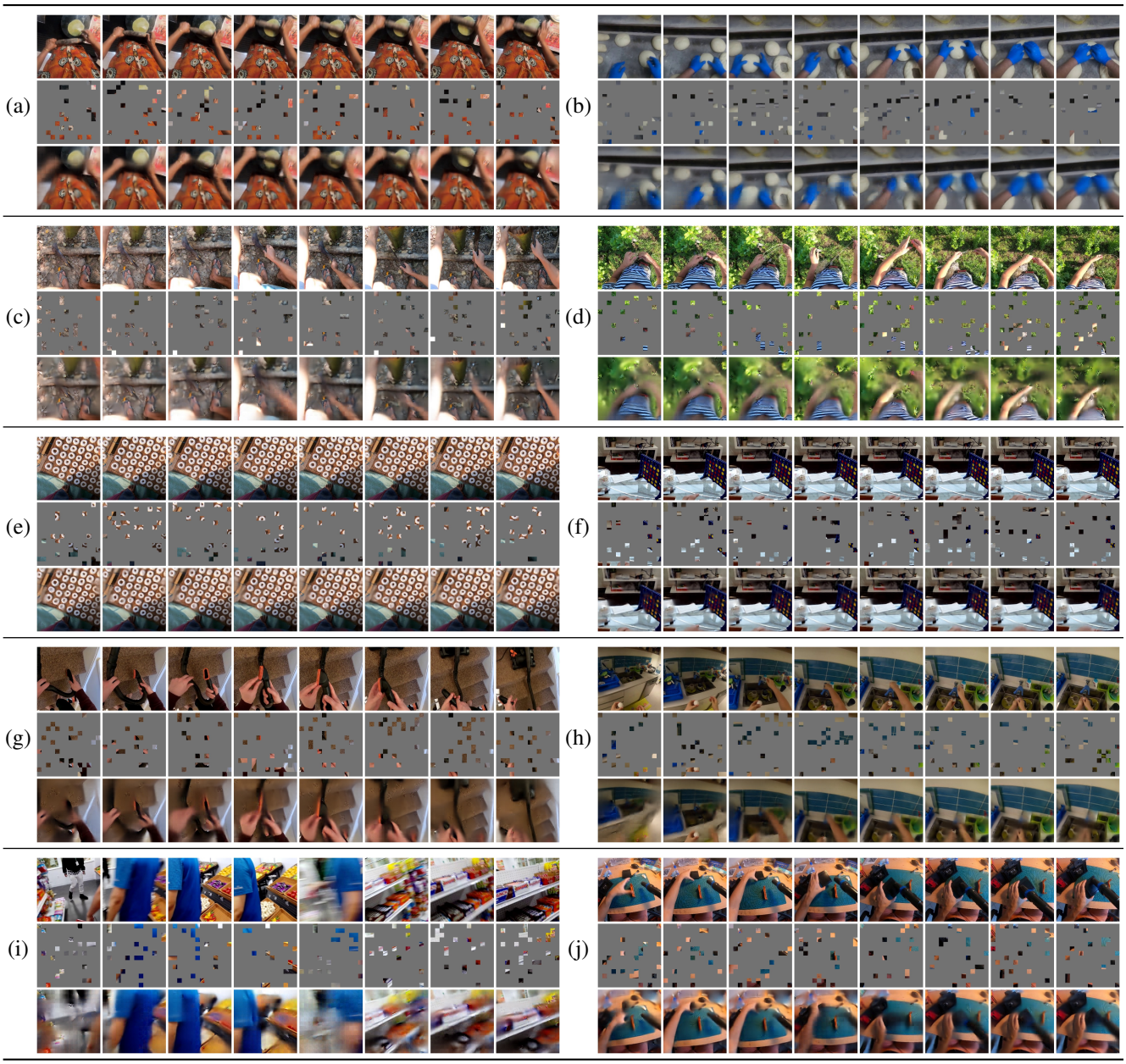

Figure 7. MAE [32, 27] reconstruction results on Ego4D [30] MQ *val* set. For each sample, we show the original video (top), the randomly masked video (middle), and the MAE reconstruction (bottom). We visualize 8 frames [27] out of 16 with a temporal stride of 2. The model predicts RGB pixels without patch normalization with a masking ratio of 90%. We notice that egocentric MAE learns human-object interactions (a,b,c,d) and temporal correspondence across frames (e,f), even in cases with strong head/camera motion (g,h,i,j).

5. Conclusion

In this work, we have shown for the first time that we can train a state-of-the-art egocentric action detector without any exocentric transferring. Our proposed Ego-Only simplifies the current learning pipeline by removing the previous need for supervised pretraining on large-scale exocentric video or image datasets before transferring to egocentric videos. We hope our such attempt inspires the community to rethink the trade-off between training in-domain with ego-only data and transferring from out-of-domain exocentric learning. We also hope that our Ego-Only results provide a strong baseline for future research that aims to improve egocentric learning by leveraging exocentric data.

Acknowledgments. We would like to thank Christoph Feichtenhofer for sharing Video MAE code and models. We thank Effrosyni Mavroudi, Gene Byrne, Mandy Toh, Triantafyllos Afouras, Yale Song, for their advice and help.

References

[1] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *ECCV*, 2018. 5

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 2, 3, 6

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 3, 6

[4] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022. 3

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 2, 3

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 1

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 3

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 3

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3, 7

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 3, 5

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 4

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017. 4

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017. 4

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 4

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[16] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 3

[17] Feng Cheng and Gedas Bertasius. Tallformer: Temporal action localization with long-memory transformer. In *ECCV*, 2022. 3

[18] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *WACV*, 2020. 5, 6

[19] Ego4D Consortium. Egocentric live 4d perception (ego4d) database: A large-scale first-person video database, supporting research in multi-modal machine perception for daily life activity. 2020. https://sites.google.com/view/ego4d/home. 1

[20] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE TPAMI*, 2020. 3

[21] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2022. 1, 2, 3, 4, 5, 6

[22] 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. 1, 5, 6, 7

[23] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 3

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2, 3, 8

[25] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2, 4

[26] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 2

[27] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 2, 3, 4, 8, 9

[28] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 2, 3, 4, 5, 6, 7

[29] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 2, 6

[30] Kristen Grauman, Michael Wray, Adriano Fragomeni, Jonathan PN Munro, Will Price, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, et al. Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 8, 9

[31] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 3

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022. 2, 3, 8, 9

[33] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[34] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, 2019. 3

[35] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2, 3

[36] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 2017. 1

[37] 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, 2, 5, 6

[38] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. In *BMVC*, 2021. 6

[39] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 3

[40] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *CVPR*, 2021. 6

[41] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7

[42] Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005. 7

[43] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*, 2021. 1, 2, 3, 5, 6

[44] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 2, 4

[45] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 2

[46] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, 2022. 3, 5, 6

[47] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, 2019. 2, 3, 6

[48] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *ACM MM*, 2017. 2

[49] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[50] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 3

[51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4

[52] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 2, 4

[53] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 7

[54] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 3

[55] Weichao Qiu and Alan Yuille. Unrealcv: Connecting computer vision to unreal engine. In *ECCV*, 2016. 1

[56] 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, 5, 6

[57] 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. 3

[58] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 8

[59] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 2, 3

[60] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[61] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 2

[62] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. MaX-Deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021. 2

[63] Jue Wang, Gedas Bertasius, Du Tran, and Lorenzo Torresani. Long-short temporal contrastive learning of video transformers. In *CVPR*, 2022. 3

[64] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE TPAMI*, 2018. 2, 4

[65] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2

[66] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *ICCV*, 2021. 6

[67] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 3

[68] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *CVPR*, 2019. 3

[69] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 6

[70] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3

[71] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, 2020. 2, 3, 6

[72] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 2, 6

[73] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *ECCV*, 2022. 2, 3, 4, 5, 7

[74] Chen Zhao, Merey Ramazanova, Mengmeng Xu, and Bernard Ghanem. Segtad: Precise temporal action detection via semantic segmentation. *ECCVW*, 2022. 3, 4

[75] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *ICCV*, 2021. 2, 3, 4, 5, 7

[76] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *arXiv preprint arXiv:2212.04501*, 2022. 6

[77] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *ICLR*, 2022. 3