SOS! Self-supervised Learning Over Sets Of Handled Objects In Egocentric Action Recognition

Victor Escorcia, Ricardo Guerrero, Xiatian Zhu, and Brais Martinez

Samsung AI Center Cambridge
{v.castillo, r.guerrero, brais.a}@samsung.com

Abstract. Learning an egocentric action recognition model from video data is challenging due to distractors in the background, *e.g.*, irrelevant objects. Further integrating object information into an action model is hence beneficial. Existing methods often leverage a generic object detector to identify and represent the objects in the scene. However, several important issues remain. Object class annotations of good quality for the target domain (dataset) are still required for learning good object representation. Moreover, previous methods deeply couple existing action models with object representations, and thus need to retrain them jointly, leading to costly and inflexible integration. To overcome both limitations, we introduce **S**elf-Supervised Learning **O**ver **S**ets (SOS), an approach to pre-train a generic Objects In Contact (OIC) representation model from video object regions detected by an off-the-shelf hand-object contact detector. Instead of augmenting object regions individually as in conventional self-supervised learning, we view the action process as a means of natural data transformations with unique spatiotemporal continuity and exploit the inherent relationships among per-video object sets. Extensive experiments on two datasets, EPIC-KITCHENS-100 and EGTEA, show that our OIC significantly boosts the performance of multiple state-of-the-art video classification models.

Keywords: handled objects, egocentric action recognition, self-supervised pre-training over sets, long-tail setup.

1 Introduction

Egocentric videos recorded by wearable cameras give a unique perspective on human behaviors and the scene context [9,32,44]. Existing action recognition methods, typically developed for third-person video understanding, focus on learning from the whole video frames [4,15,28,6,34,51,60]. Unlike third-person videos where actions often take place in dramatically different background scenes (e.g., swimming in a pool, cycling on the road) [6,45], egocentric videos are often collected at a specific scene (e.g., a kitchen) with similar background shared across different human actions (e.g., cutting onion and washing knife) and cluttered

Fig. 1. Handled Objects are vital for determining actions in egocentric videos. However, most action classification models learn directly from video frames (left). This paper puts manipulated objects (right) at the forefront *without* the need for expensive and tedious fine-grained object annotations.

with distractors (e.g., a knife on the countertop). These distinctive characteristics bring extra challenges for most existing action models *without a fine-grained understanding of spatiotemporal dynamics and context.*

Recent works have incorporated information from other modalities, such as audio [28,29], narration language [3,27], and eye gaze [31], to overcome these challenges in the video action model. This is motivated by their complementary information and the modality-specific nature of distractors. Object information is a powerful complement to prior video action models [12,13,52,56]. While a video model represents the sequence as a whole, combining foreground and context, object detectors exploit bounding box annotations to explicitly model each object, separately from the others and background.

Our work falls in this line of research with a crucial difference. We consider that bounding box annotation is costly and unsuited for *long-tail* open-vocabulary object distributions across the scenes, e.g., a kitchen setting. Large-scale annotations are thus seldom possible in most cases. We tackle this issue by capitalizing an *off-the-shelf hand-object contact detector* to localize class-agnostic object regions in the video. Crucially, we introduce a novel self-supervised approach to learning a specialized representation of the detected object regions without resorting to object-level labels. Our key idea is to leverage the spatiotemporal continuity in videos as *a native context constraint* for exploiting the inherent relationships among the set of detected object regions per video. Intuitively, all the handled object regions play respective roles during an action, and they collectively provide a potentially useful context clue for learning a suitable object representation (e.g., the "pan" and "spoon" while mixing - see Fig. 1). With our class agnostic self-supervised learning, there is also a potential that the natural action class imbalance problem would be simultaneously alleviated – a typical yet understudied problem in video understanding. We term our self-supervised pre-training **S**elf-Supervised Learning **O**ver **S**ets (SOS). After pre-training, we transfer the specialized representation of detected objects to a target task, e.g., video classification. For this purpose, we employ an Objects in Contact (OIC) network and further fine-tune the entire representation with *weak* video-level labels. Once trained, our OIC can be flexibly integrated with existing video classification models, further boosting their performance.

We make three **contributions** in this paper. **(1)** We investigate the merits of learning a representation of handled objects for egocentric action recognition, easy to integrate over multiple state-of-the-art video action models without the

need for expensive fine-grained region-level labeling on the target dataset. **(2)** To that end, we leverage an off-the-shelf hand-object contact detector to generate class-agnostic object regions and introduce a novel set self-supervised learning approach, SOS, to learn a specialized representation of object regions. SOS exploits the inherent relation (e.g., temporal continuity and concurrence) among handled objects per video to mine the underlying action context clue by treating all the objects collectively as a set. **(3)** Experiments on two egocentric video datasets showcase that our OIC, aided with our SOS pre-training, complements multiple existing video classification networks and yields state-of-the-art results in video classification. Moreover, we demonstrate the benefit of SOS for dealing with the realistic long-tail setup in videos. (Sec. 4.3).

2 Related Work

Egocentric action recognition. Egocentric action recognition has made significant advances in recent years, thanks to the introduction of ever-larger video benchmarks [9,10,32]. Early efforts were focused on adapting representative generic video models [15,34,51,60]. Later on, a variety of dimensions were investigated. For example, Kazakos et al. [28] combined multi-modal information (e.g., optical flow, audio) within a range of temporal offsets. Bertasius et al. [3] leveraged the language-based semantic context to supervise the learning of active object detection. Li et al. [30] investigated the pre-training of a video encoder for mitigating the domain gap from the common pre-trained video datasets (e.g., Kinetics [6]). Sudhakaran et al. [46] designed an LSTM-based Long-Short Term Attention model to locate relevant spatial parts (e.g., active objects) with temporally smooth attention. Similarly, Yan et al. [57] proposed an Interactive Prototype Learning (IPL) model for better learning active object representations by interaction with different motion patterns. Instead, Li et al. [32,31] and Liu et al. [35] used human gaze to guide the attention of deep models to interacting regions. Similar to ours, object detection has been previously exploited to improve action recognition. Indeed, early work already identified explicit hand detection as an informative queue for action recognition [1]. More recently, Wang et al. [54] exploited object regions and their spatiotemporal relations to enhance video representation learning. Similar to ours, Baradel et al. [2] devised an architecture with a video branch and an object branch. However, their approach requires object-level annotations and has no mechanism to identify foreground/active objects, thus being vulnerable to distractors. Wu et al. [56] used an attentive mechanism and incorporated long-term temporal context. Wang et al. [52] proposed a model for egocentric action recognition that relies on a complex attentive mechanism to sieve out distractors. Unlike prior work, our method does not assume fine-grained object-level annotations from the target dataset, hence being scalable in practical applications. Critically, their object representation is tightly coupled with an action model with the need for joint training. In contrast, we learn an independent object representation model enabling to flexibly benefit from off-the-shelf action models in a decoupled post-training manner.

Hand-object interaction (HOI). While generic Human-Object Interaction is a widely-studied topic [16,40,33,22], recent works have shown that it is possible to train large-scale domain-agnostic hand and hand-object contact detectors [43,37]. As face detection models, these models can be deployed without domain-specific retraining and still maintain reasonable effectiveness. We apply one such off-the-shelf hand-object contact model without retraining in our work [43].

Self-supervised learning. There has been a recent surge in self-supervised learning (SSL) for learning generic feature representation models from large-scale datasets without labels [20,8,7,5,18,48]. Inspired by this trend, we introduce a novel application of exiting SSL techniques to learn a representation specific for handled objects. A direct application of an SSL algorithm to our problem does not produce optimal results. We propose two important modifications: firstly, instead of following a classic fine-tuning from domain-agnostic pre-training, e.g. over ImageNet, our method leverages on-domain SSL for solving the domain shift problem with model pre-training. To this end we use a domain-agnostic SSL model as means of pre-training, followed by domain-specific SSL training [41]. Secondly, existing SSL methods focus on single images, using two different augmentations to obtain two copies of an image. Instead, in the presence of video, sampling different timestamps and locations can create more natural and effective augmentations [39]. Inspired by this insight, we propose a variant of SwAV that operates on sets of image regions extracted from a video sequence.

Long-tail learning in video. Real-world egocentric actions are typically class-imbalanced [9,10,32]. Despite extensive works in image domains [11,23,25,26,36,58], class imbalance is still less studied in video tasks [59]. Inspired by the intriguing finding that SSL learns a suitable representation with class-imbalance scenarios [58], we evaluate if the insights also apply to videos. Unlike [58] images, video data is more complex due to the extra temporal dimension and the structured nature of action labels (*i.e.*, defined as a tuple of two imbalanced label distributions, verb, and noun) and only access to video-level supervision. This paper contributes to the first study of the relationship between SSL and long-tailed learning in egocentric video understanding.

3 Method

We aim to learn a generic object-in-contact (OIC) model for improving the performance of existing egocentric action models in a plug-and-play manner.

Overview. Given a video V and an off-the-shelf hand-object contact detector model [43], we obtain a set of object regions that likely contain objects manipulated by the hands. Let $\mathcal{B} = \{\boldsymbol{B}_i\}_{i=1:M}$ denote the set of M object regions in the video. We have an object region encoder f_B so that $\boldsymbol{y}_B^i = f_B(V, \boldsymbol{B}_i; \theta_B)$. To that end, we propose a simple yet effective approach termed, **S**elf-Supervised Learning **O**ver **S**ets (SOS), to train θ_B in two steps: **(I)** First, a large-scale self-supervised learning model is used for pre-training. **(II)** Followed by, an on-domain self-supervised learning stage yielding a specialized representation better suited for the task of interest. Given a target task, standard discriminative fine-tuning is

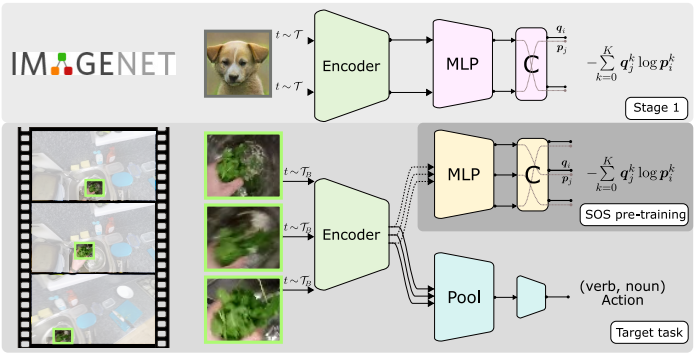

Fig. 2. Overview of the proposed Self-Supervised Learning Over Sets (SOS) approach for pre-training a specialized Objects In Contact (OIC) representation model. Taking as input object regions extracted by an off-the-shelf hand-object contact detector in videos, we formulate a two-staged pre-training strategy. The *first* stage consists of generic self-supervision learning (e.g. ImageNet). In the *second* stage, we exploit self-supervised learning on per-video object sets using actions as the natural transformations with spatiotemporal continuity. Given a target task, we further fine-tune the representation through *video-level* action labels.

followed, using the video-level labels and standard cross-entropy as supervision. An overview of our SOS is depicted in Fig. 2.

3.1 Self-Supervised Object Representation Learning from Video Object Regions

Object representations are typically learned as part of the standard object detector training pipeline [52]. However, we do not assume the availability of object-level annotations. Instead, we learn the object representation in an unsupervised manner using class-agnostic regions from the hand-object contact detector [43].

Stage I: Model pre-training. There is a lack of standard large-scale datasets for model pre-training in the egocentric video so that one could use the widely used ImageNet [42] supervised pre-trained weights for model initialization. This gives rise to an inevitable domain shift challenge for object representation learning due to the intrinsic discrepancy in data distribution. To overcome these domain shift challenges, we leverage the more domain-generic self-supervised learning (SSL) strategy for model pre-training [7,18,20,50,8,5]. In practice, due to the relatively small size of the target dataset, it is key to start with an SSL model trained on ImageNet for model initialization. In general, any existing SSL method is applicable. Based on preliminary experiments, we select the recent state-of-the-art model SwAV [5] (Fig. 3 top - Stage I).

Stage II: On-domain Self-supervised learning Over Sets, SOS. The key challenge is how to capitalize the unlabeled object regions. Motivated by the results of [41], we perform on-domain SSL to specialize the object region encoder f_B to better represent the regions in the target domain, egocentric videos.

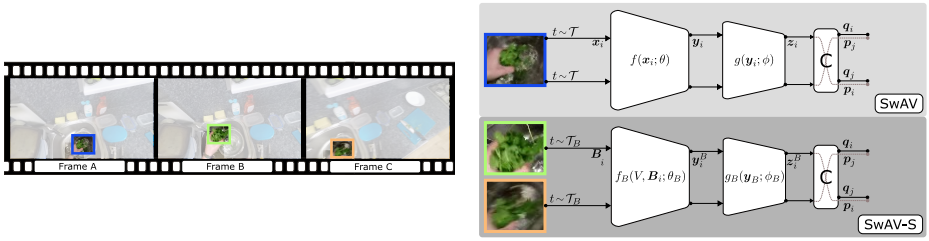

Fig. 3. Overview of the proposed SwAV-S. Schematically outlining the differences w.r.t.the standard SwAV [5] for image regions, and our self-supervised pre-training approach using videos and leveraging actions as natural transformations.

The promising pre-training recipe from [41] ignores a crucial piece of information from our problem definition, object regions extracted from a single video are *not* independent but *collectively* correlated due to the ongoing action. Inspired by this consideration, we introduce the notion of actions as natural data transformation exploiting the unique spatiotemporal continuity of videos and the inherent relations among a collection of objects per video. We argue that statistically predominant inter-object relationships (*e.g.*, temporal continuity and concurrence) provide a strong self-supervisory signal for learning a specialized object representation suitable for action recognition.

We consider each video V as a training sample and treat the *set construction process from all regions in the video* $\mathcal{B}$ as *a spatiotemporal transformation process*, subject to the structural variations of the performed action in the video (*i.e.*, elastic, geometrical or ambient changes that objects undergo through space and time). This is conceptually reminiscent of and complements the standard image augmentation process (e.g., cropping, flipping, jittering). Interestingly, under this perspective, our approach fits exactly on recent top-of-the-line self-supervised frameworks such as contrasting learning, clustering, instance similarity, and decorrelated representation, to name a few [7,18,20,8,5]. Without loss of generality, we present our Self-supervised Learning Over Sets approach (SOS) on top of the SwAV formulation due to their state-of-the-art results. Yet, we anticipate that our SOS could be extended to other frameworks[7,18,20,8].

SwAV review. SwAV [5] aims to learn a feature representation that matches different views of the same image. Specifically, the representation of one of these views is used to compute a code $\boldsymbol{q}$, which is then predicted from the code of another view. Formally, given an image $\boldsymbol{x}$, two corresponding views $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are created by applying a random transformation $t \in \mathcal{T}$, where $\mathcal{T}$ is the set of all considered transformations (typically these are synthetic, such as random crops, color jittering, Gaussian blurring, and flipping, etc.). Corresponding feature vectors $\boldsymbol{z}_i = g(f(\boldsymbol{x}_i;\theta);\phi)$ are generated and projected into the unit sphere, where $\boldsymbol{z}_i \in \mathbb{R}^d$, f represents a backbone network with parameters θ and g a projection head with parameters ϕ. SwAV can be seen as an online clustering method, where the cluster centroids are defined by a set of learnable prototypes $\boldsymbol{c}_i \in \boldsymbol{C}^{K\text{x}d}$, which are used as a linear mapping function to compute each view's

code $\boldsymbol{q}_i = \boldsymbol{z}_i^\top \boldsymbol{C}$. The objective is to predict a view's code $\boldsymbol{q}_i$ from the other view's features $\boldsymbol{z}_j$, and the problem is formulated via cross-entropy minimization as:

$$L(\boldsymbol{z}_1, \boldsymbol{z}_2) = \ell(\boldsymbol{q}_1, \boldsymbol{z}_2) + \ell(\boldsymbol{q}_2, \boldsymbol{z}_1) \tag{1}$$

where the first term is

$$\ell(\boldsymbol{q}_1, \boldsymbol{z}_2) = -\sum_{k=0}^{K} \boldsymbol{q}_1^k \log \boldsymbol{p}_2^k \ , \tag{2}$$

where $\boldsymbol{q}_1$ represents the prototype likelihood or soft cluster assignment of $\boldsymbol{x}_1$, $\boldsymbol{p}_2^k$ represents the prediction of $\boldsymbol{q}_1$ as the softmax of $\boldsymbol{z}_2^\top \boldsymbol{c}_k / \tau$ and τ is a temperature parameter. Additionally, $\boldsymbol{q}_1^k$ comes from $\boldsymbol{Q}$, which normalizes all $\boldsymbol{q}$ in a mini-batch (or queue) using the Sinkhorn-Knopp algorithm. The second term of Eq. (1) is similarly defined. In practice, SwAV uses more than two views per image. More specifically, the concept of multi-crop is introduced, where crops taken as part of the augmentation strategy can be either global or local concepts of an image. Furthermore, the problem is formulated in such a way that prototype code $\boldsymbol{q}$ assignments are only done on global views, while predictions $\boldsymbol{p}$ are done using both local and global views.

Our SwAV-S. To better exploit unlabeled object regions from videos, we introduce a set structure into SwAV's formulation, resulting in a new SSL variant dubbed as *SwAV-S*. Specifically, we sample a subset of object regions from $\mathcal{B}^{'} = \{\boldsymbol{B}_i\}_{i=1:N}, \mathcal{B}^{'} \subset \mathcal{B}$ from each video. Then, each region undergoes an independent image transformation and is treated as one view of V to be predicted (or contrasted) from another region of the same set. This set of regions are encoded, generating embedding vectors $\{\boldsymbol{z}_i = g_B(f_B(V, \boldsymbol{B}_i; \theta_B); \phi_B)\}_{i=1:N}$, where f_B and g_B represent a non-linear encoder function and projection head, respectively.

In contrastive learning design, we make a couple of important differences against the original SwAV. (1) At each training iteration, we sample $N > 2$ object regions $\{B_i\}_{i=1:N}$ from a video sequence V. (2) We only consider global views, which allows the expansion of terms in Eq. (1) to N, effectively treating all object regions as a set as follows:

$$L = \frac{-1}{N^2 - N} \sum_{i}^{N} \sum_{j \neq i}^{N} \sum_{k=0}^{K} \boldsymbol{q}_i^k \log \left(\frac{\exp\left(\frac{1}{\tau} \boldsymbol{z}_j^\top \boldsymbol{c}_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau} \boldsymbol{z}_j^\top \boldsymbol{c}_{k'}\right)} \right) \tag{3}$$

where (i, j) indexes the pairs of regions from a video.

Discussion. While our approach shares some high-level ideas w.r.t. earlier SSL algorithms [55], SOS relies on (1) different assumptions (*e.g.*, do not require an explicit graph of relations among patches or tracking), (2) different aims (integrating object representations for action recognition), and (3) specific methodology (self-supervised learning over sets). All in all, our approach revives this line of research with a refreshing perspective and using leveraging recent insights.

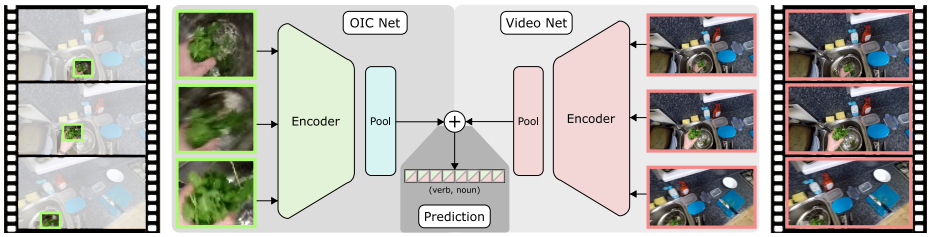

Fig. 4. Our OIC model can complement any existing video classification network, Video Net. While most existing classification networks consume the entire video frames, ours OIC only takes object regions. Their predictions are then late fused.

3.2 Target task fine tuning: OIC Net and Model Fusion

Following the typical transfer learning of SSL pipelines, The pre-trained enhanced object region encoder θ_B serves as initialization for the OIC encoder f_B addressing the target task (*cf.* Fig. 4 - OIC Net). In our case, the target task corresponds to supervised video classification using object patches.

Given a set of object regions $\mathcal{B}$ per video, we feed them through the OIC encoder, followed by a pooling module aggregating the information from all the cohort to obtain the classification logits of interest $l = h(f_B(V, \mathcal{B}; \theta)) \in \mathbb{R}^{|\mathcal{Y}|}$, where $|\mathcal{Y}|$ corresponds to the size of the label space. Note that we only have access to weak video-level supervision. We apply the standard cross-entropy loss between the logits and the video-level label to optimize the network parameters.

Class imbalanced fine-tuning. To complement our SOS pre-training for tackling the class imbalance issue, we adopt the recent logit adjustment method [36] with the key idea is to impose the class distribution prior of the training data through logit regulation. For training, we optimize a logit adjustment cross-entropy loss, $L_{la} = -\log \frac{e^{f_y + \tau \log \pi_y}}{\sum_{y' \in \mathcal{Y}} e^{f_{y'} + \tau \log \pi_{y'}}}$ with π_y the frequency of class y on the training set (i.e., the class prior), and τ the temperature. In our case, we impose this adjustment to the noun and verb branches separately, so their imbalance can be remedied according to their respective distribution priors.

Model Fusion. Once trained as discussed above, our OIC model can be integrated with any existing action models [4,34,15] as depicted in Fig. 4. For simplicity and flexibility, we use a weighted late-fusion of video predictions as:

$$\boldsymbol{l} = \alpha_{\text{OIC}} \boldsymbol{l}_{\text{OIC}} + \alpha_i \boldsymbol{l}_i, \tag{4}$$

where $\boldsymbol{l}_{\text{OCI}}, \boldsymbol{l}_i$ are the classification logits of our OIC and any existing model; and $\alpha_{\text{OCI}}, \alpha_i$ are scalars weighting the confidence of each model. In practice, we set $\alpha_{\text{OCI}}, \alpha_i$ based on the performance of each model on a pilot validation set (*e.g.*, 30% of the training set). The fusion is applied separately in the case of verb and noun-based action prediction.

4 Experiments

Datasets. We evaluate our method on two standard egocentric video datasets in the domain of kitchen environments. ***EPIC-KITCHENS-100*** [10] is a large-scale egocentric action recognition dataset with more than 90,000 action video clips extracted from 100 hours of videos in kitchen environments. It captures a wide variety of daily activities in kitchens. The dataset is labeled using 97 verb, and 300 noun classes and their combinations define different action classes. Please refer to the **supplementary material** for results in the test server, and additional details, among others. ***EGTEA*** [32] is another popular egocentric video dataset consisting of 10,321 video clips annotated with 19 verb classes, 51 noun classes, and 106 action classes.

Performance metrics. We adhere to the standard action recognition protocol [10,32,53]. Specifically, each model predicts the verb and the noun using two classification heads, and we report accuracy rates of verb, noun, and action (*i.e.* verb and noun tuple) as performance metric. For EPIC-KITCHENS-100, we report the top-k accuracy over different sets of instances. For EGTEA, we adopt the mean class accuracy on the three train/test splits.

Implementation details. We use an off-the-shelf hand-object contact detector to extract candidate regions of interest where it is likely to find hands and manipulated objects [43]. We only consider the object regions and select those with a confidence threshold greater than 0.01 non-filtered by the typical non-maxima suppression operation of object detector pipelines [16]. Each region of interest is cropped from the original frame size and resized such that the resolution of each object region is 112×112. During training, we enable the center and size jittering as data augmentation. The jittering is proportional to the size of the region of interest and sampled from a uniform distribution $[1, 1.25)$. For each frame, we consider three regions at most. At test time, we retained the most confident detected regions. Yet, we sampled object regions irrespective of the confidence score during training for data augmentation purposes.

Our OIC network uses a 2D-CNN backbone, ResNet-50 [21], to encode each object region. Its Pool block corresponds to an integrated classifier followed by a aggregation module, the video prediction is done by classifying each object region independently, and then aggregating all the object predictions with a parameter-less Mean pooling operation. We use the standard frame sampling [51], and modestly consider 8 frames per video. In contrast to prior art [4], we did not resort on additional test-time data augmentations (*e.g.* multi-crop/views) per video. Thus, the performance of our model could improve further.

During the on-domain SOS self-supervised pre-training (Stage II), we sample sets of size $N = 8$ objects per video, and apply the standard set of photometric and geometric data augmentation (*e.g.*, color jittering and cropping *cf.* [5] for more details) independently per element. We resort to single-machine training with a batch size of 256 over 400 epochs and used the default optimization hyperparameters [5]. We initialize the CNN backbone of OIC from an off-the-shelf self-supervised pre-trained network on ImageNet [5] (*i.e.*, Stage I pre-training). During the target task stage, we tune the whole OIC network end-to-end for

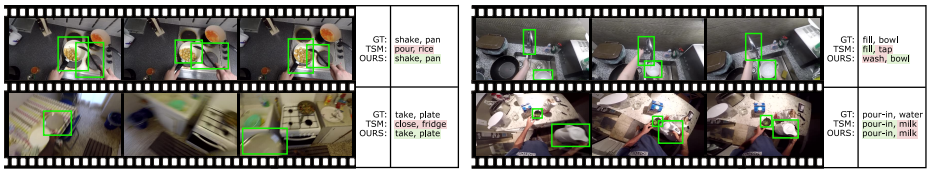

Fig. 5. Qualitative results. The green bounding boxes represent automatically detected object regions being manipulated. The left examples are corrected by our method, while the right two examples either error were maintained or flipped.

video classification. We train with SGD using an initial learning rate of 0.02 and momentum 0.9 with a batch size of 64 videos. We decay the learning rate by a factor of 0.1 after 20 epochs. We implemented our approach in Pytorch [38], PytorchLightning [14] and the NUMFocus stack [19,49].

4.1 Evaluation on EPIC-KITCHENS-100

Qualitative analysis Fig. 5 depicts qualitative results, each quadrant depicts three frames from a particular video with its associated object regions, ground truth labels, and the predictions made by TSM as well as OURS (*i.e.*, TSM+OIC). Correctly predicted verbs and nouns are highlighted as green, while incorrectly predicted as red. We observe that OIC network provides additional context to guide predictions correctly. Fig. 5 (bottom left) depicts an example where the noun of interest is mainly out of view during the duration of clip (only visible at the very beginning and at the end), hence the verb is also implicitly also not visible. Nevertheless, including the OIC module allows this information to be correctly captured and preserved. Similarly, Fig. 5 (top right) shows that TSM alone makes somewhat nonsensical prediction of "fill, tap", while ours predicts "wash, bowl" which is arguably visually and semantically closer to the ground truth "fill, bowl", after all filling a bowl looks much like washing a bowl. However, both TSM and ours still struggle with visually confusing noun classes, as shown in Fig. 5 (bottom right). Although both models correctly predict the verb "pour-in", they both confuse the white water kettle with a jug of milk. Perhaps surprisingly, we appreciate that TSM often struggles with seemingly simple cases (noun and verb clearly visible and executed), as shown in Fig. 5 (top left), while the inclusion of OIC greatly alleviates this problem.

Baselines For extensive comparative evaluation, we consider a variety of state-of-the-art action recognition models including (1) top CNN action models: TSM [34], and SlowFast [15] designed for generic action recognition tasks with varying architectural design for spatio-temporal representation learning, (2) a recent vision transformer: X-ViT [4] with superior cost-effective formulation for spatio-temporal attention learning, and (3) the latest egocentric action model: IPL [53] designed to learn superior active object representation subject to human motion cues. We test the effective of our OIC in improving several above methods with the proposed fusion method.

Modality	Method	TTDA	Overall						Unseen participants			Tail classes		
			Top-1			Top-5			Top-1			Top-1		
			Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
V	TSM† [34]	✗	63.0	47.3	35.6	88.9	74.2	57.2	**54.6**	37.1	26.3	35.9	26.7	18.1
V	+ **OIC**	✗	**64.0**	**52.3**	**39.0**	**90.2**	**78.4**	**61.6**	53.1	**40.3**	**27.2**	**36.2**	**29.9**	**20.5**
V	SlowFast† [15]	✗	65.6	50.0	38.5	**90.0**	75.6	58.7	**56.7**	41.6	30.0	36.3	23.3	18.7
V	+**OIC**	✗	**66.5**	**53.5**	**40.9**	89.6	**77.8**	**61.8**	56.6	**43.6**	**31.1**	**37.6**	**27.3**	**20.3**
V	X-ViT† [4]	✗	62.9	49.1	37.7	87.5	73.8	57.1	54.4	41.7	30.5	35.7	28.0	19.42
V	+**OIC**	✗	**64.7**	**54.6**	**40.9**	**91.0**	**79.3**	**62.8**	**54.6**	**43.8**	**31.2**	**37.0**	**31.1**	**21.2**
V	X-ViT [4]†	✓	68.7	56.4	44.3	90.4	79.4	64.5	57.8	47.4	34.9	38.2	31.8	22.4
V	+**OIC**	✓	**69.3**	**57.9**	**45.7**	**91.2**	**81.1**	**66.2**	**57.9**	**49.3**	**36.3**	**38.6**	**33.1**	**23.7**
V+F	TSM† [34]	✗	67.9	49.1	38.4	91.0	74.9	60.7	58.5	39.2	29.3	36.8	22.5	17.7
V+F	+**OIC**	✗	**68.4**	**53.7**	**41.8**	**91.4**	**78.7**	**64.0**	**58.7**	**42.7**	**30.8**	**37.1**	**26.7**	**20.1**
V+F	IPL(I3D) [53]	✗	67.82	50.87	39.87	-	-	-	-	-	-	-	-	-
V+F	IPL(R2+1D) [53]	✗	68.61	51.24	40.98	-	-	-	-	-	-	-	-	-

Table 1. Results on the validation set of EPIC-KITCHENS-100 [10]. Modality: V=Visual; F=Optical flow. †: Results computed with the model weights released by the authors of [10,4]. TTDA: Test-time data-augmentation (*e.g.*, multi-crop). Underlined numbers correspond to best results across the board. Bold numbers highlight the best between a proxy model and proxy + ours. All in all, our model yields state-of-the-art results and it is versatile as improved four strong action classification models.

Results analysis We report the action recognition results in Table 1. We have the following observations: **(1)** The performance of CNN action models, SlowFast and TSM, are similar to the recently introduced X-ViT without heavy test-time data augmentation. **(2)** Importantly, our OIC further improves SlowFast(V), TSM(V+F) and X-ViT by 2.4%, 3.4%, and 3.2% on overall action accuracy, respectively. Without using computationally expensive optical flow, TSM still benefits from OIC at a similar scale. These results verify the versatility and consistent usefulness of our OIC in enhancing prior art models. As expected, the improvement is mainly achieved in noun recognition. This evidence indicates that generic action recognition methods are limited in modeling the handled objects from cluttered backgrounds and scenes. This is exactly the motivation for learning our OIC model. **(3)** We also observe that our OIC clearly improves the accuracy scores on unseen participants and tail classes. This implies that exploiting our OIC could help reduce the negative impact of domain shift (seen vs. unseen participants in this case) and mitigate the overwhelming effect from head classes to tail classes, concurrently. **(4)** Along with TSM, our method also clearly outperforms the latest egocentric action model IPL designed to learn better active object representations. This indicates that without explicit region detection, the CNN model is less effective in isolating the active objects from the scenes. **(5)** The recent X-ViT model lifts the performance of CNN using additional test-time data augmentation (*i.e.*, averaging the results from 3 crops per video). It is worth noting that even after triplicating the computational budget, our computationally modest OIC representation stills provides a further gain of 1.4% on overall actions to this model. ***Computational complexity and runtime***. In a 1080Ti GPU, TSM and XViT run in 12 and 66 msec. respectively,

Modality	Method	Split 1	Split 2	Split 3	Avg
V	TSM [34]	61.2	61.2	59.6	60.7
V	TSM+**OIC**	**62.0**	**61.9**	**60.0**	**61.3**
V+F	I3D* [6]	55.8	53.1	53.6	54.2
V+F	Ego-RNN-2S* [47]	52.4	50.1	49.1	50.5
V+F	LSTA-2S* [46]	53.0	-	-	-
V+F	MutualCtx-2S* [24]	55.7	-	-	-
V+F	Prob-ATT [31]	56.5	53.5	53.6	54.5
V+F	I3D+IPL [53]	60.2	59.0	57.9	59.1
V+F+G	I3D* [6]	53.7	50.3	49.6	51.2
V+F+G	Prob-ATT [31]	57.2	53.8	54.1	55.0

Table 2. Comparison against state of the art in EGTEA [32]. Modality: V=Visual; F=Optical flow; G=Gaze. *: Results are taken from [31].

while the OIC module runs in 10.7 msec. The OIC complexity is 26 GFLOPs while XViT is much larger, 285 GFLOPs. OIC is comparatively efficient.

4.2 Evaluation on EGTEA

Baselines Compared to EPIC-KITCHENS-100, EGTEA uniquely features eye gaze tracking information. We compare our method with the following alternatives: (1) I3D [6]: A two-stream I3D with joint training of RGB and optical flow, (2) I3D+Gaze: Instead of average pooling, the ground truth human gaze is leveraged to pool the features from the last conv layer, (3) Ego-RNN-2S [47] and LSTA-2S [46]: two recurrent networks with soft attention, (4) MutualCtx-2S [24]: A gaze-enhanced action model trained by alternating between gaze estimation and action recognition, (5) Prob-ATT [31]: A state-of-the-art joint gaze estimation and action recognition model featured with a stochastic gaze distribution formulation, (6) TSM [34]: A recent strong action recognition model with efficient temporal shift operation between nearby frames for motion modeling, (7) IPL [53]: A recent state-of-the-art egocentric action model as discussed earlier. We combine our OIC with TSM in evaluating this dataset.

Results analysis. Table 2 reports the results. We observed that: **(1)** using gaze modality does not guarantee superior results; For example, without gaze IPL still achieves better results than Prob-ATT using gaze. This suggests that video data alone already provide rich information and the key is how to learn and extract discriminative action information with proper model design. The way to leverage the gaze data is also equally critical. **(2)** Similarly, the computationally expensive optical flow is not the highest performance promise, as indicated by the excellent performance with TSM using only 2D video frames as input. **(3)** Importantly, our OIC again further improves the performance of TSM consistently over all the splits, suggesting the generic efficacy of our method on a second test scenario.

4.3 SOS and Long-tail Learning

Here we present a rigorous assessment on the role of our SOS pre-training for dealing with class-imbalance distributions in videos. ***Dataset and metrics***.

	Method	SOS	LT loss [36]	Overall			Tail classes		
				Verb	Noun	Action	Verb	Noun	Action
A	**OIC**	✗	✗	18.2	21.1	8.7	13.2	11.9	6.1
B	**OIC**	✗	✓	27.5	25.0	8.1	25.4	19.9	7.1
C	**OIC**	✓	✗	20.8	25.0	**10.8**	15.9	16.7	8.6
D	**OIC**	✓	✓	**30.3**	**28.0**	9.4	**28.6**	**23.4**	**8.7**
E	X-ViT	✗	✗	22.1	25.9	12.8	15.9	17.8	9.5
F	X-ViT + **OIC**	✓	✓	**30.2**	**31.8**	**15.2**	**25.2**	**24.9**	**12.8**

Table 3. Long-tail results (accuracy) on the validation set of EPIC-KITCHENS-100, showcasing the impact of SOS for dealing with class-imbalance distributions. Refer to text for details about the metric.

We perform the study in the validation set of EPIC-KITCHENS-100 using the standard metrics for evaluating long-tail learning algorithms [11], class-balanced average accuracy for verb, noun and action. Concretely, each video is weighted by the inverse of the number of instances of its corresponding label. Note that we retain the tail classes definition from [10], but the metrics reported in this section differ from those in Table 1. ***Baselines***. We consider our OIC network aided (or not) with our SOS pre-training and using (or not) the long-tail (LT) loss of [36] during the target task fine-tuning. We also report of single-crop X-ViT [4], using the pre-trained weights of the authors, and our fused model.

Results analysis. Table 3 reports the results. We observed that: **(1)** Our SOS pre-training indeed helps for dealing with class-imbalanced scenarios. It naturally improves noun and action classes the most, by +3.9% and +2.1% overall classes respectively (row **C** v.s. **A**). Note that aiding OIC with SOS (**C**) significantly reduced the gap between the X-ViT (**E**) and our plain OIC (**A**). **(2)** SOS is complementary to the state-of-the-art LT learning approach. SOS and the LT loss [36] yields the best results for the OIC model, except for overall action classes where SOS pre-training achieves the best result. The relatively minor setback evidences the relevance of studying LT and SSL pre-training within the context of structured labels such as actions in videos. **(3)** Our fused model (**F**) , X-ViT with OIC aided with SOS and the LT loss [36], achieves the state-of-the-art on LT accuracy for noun and action by a large margin.

Overall, we have validated the relevance and complementary of SOS for dealing with long-tail learning scenarios in egocentric action recognition.

4.4 Ablation Study

As prior art [53], we validate the major components of our approach on EPIC-KITCHENS-100 using the corresponding evaluation protocol [10].

Off-the-shelf self-supervised encoder. We study the impact of using off-the-shelf supervised [38] vs. self-supervised [5] encoders for representing handled objects in the target dataset. Both encoders were trained in ImageNet [42], and serve as initialization for our OIC backbone. Table 4 row **A** and **B** report the

	Method	OD	Overall			Tail classes		
			Verb	Noun	Action	Verb	Noun	Action
A	Supervised pre-training	✗	49.3	44.5	27.4	32.0	23.6	14.8
B	SwAV	✗	49.2	45.8	27.9	31.2	21.2	14.0
C	SwAV	✓	49.9	47.0	28.7	30.6	23.1	14.6
D	SwAV-S	✓	**51.5**	**48.5**	**30.2**	**32.4**	**25.6**	**16.4**

Table 4. Ablation of different components of our approach on EPIC-KITCHENS-100. Rows **A-D** report the top-1 accuracy Verb, Noun, and Action of the OIC model by itself (*i.e.*, without video network fusion) for *overall* and *tail* instances [10]. OD: On-Domain pre-training in target dataset, EPIC-KITCHENS-100.

results of supervised and self-supervised encoders, respectively. Self-supervised initialization improves the performance w.r.t. supervised initialization in noun by 1.3% and action by 0.5% without degrading verb performance. These results echo the relevance and popularity of self-supervised pre-training [17] in the domain of egocentric visual perception, which has been relatively under-explored.

Impact of self-supervised adaptation. We gauge the impact of adopting the off-the-shelf self-supervised representation to the domain of interest, manipulated object regions in EPIC-KITCHENS-100. For this purpose, we employ the standard SwAV loss (Eq. (1)) during the Stage II pre-training, (*cf.* Fig. 2). We use the original set of data augmentations and optimization hyper-parameters as in [5]. Table 4 row **C** reports the results of the self-supervised domain adaptation with SwAV loss over individual object regions (Fig. 3 - top). Adapting the representation helps to boost the predictive power for verb, noun and action by 0.7%, 1.2% and 0.8% w.r.t. no domain adaptation Table 4 row **B**.

Relevance of SOS. We validate the impact of SwAV-S (*i.e.*, our incarnation of SOS), which treats actions as a natural self-supervised transformation. For this purpose, we employ the SwAV-S loss (Eq. (3)) during the Stage II pre-training (see Fig. 2). We kept the same set of hyper-parameters as SwAV. Table 4 row **D** reports the results of SwAV-S. We observe the best performance across the board and a more significant boost w.r.t. on-domain SwAV (Table 4 row **C**) by 1.6% for verb, 1.5% for noun and 1.5% for action. This result validates the benefit of our novel approach for self-supervised domain adaptation.

5 Conclusions

We presented a novel approach, Self-supervised Learning Over Sets (SOS), for learning an object representation suitable for egocentric action recognition without needing object-level annotations. SOS exploits the temporal consistency and concurrence relations among a set of handled objects as a self-supervisory signal. Experiments show that prior state-of-the-art video models consistently benefit from our object representation, with improved ability to tackle the challenging long-tail setup.

References

1. Bambach, S., Lee, S., Crandall, D.J., Yu, C.: Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In: ICCV (2015)
2. Baradel, F., Neverova, N., Wolf, C., Mille, J., Mori, G.: Object level visual reasoning in videos. In: ECCV (2018)
3. Bertasius, G., Torresani, L.: Cobe: Contextualized object embeddings from narrated instructional video. In: NeurIPS (2020)
4. Bulat, A., Perez-Rua, J.M., Sudhakaran, S., Martinez, B., Tzimiropoulos, G.: Space-time mixing attention for video transformer. In: NeurIPS (2021)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
8. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021)
9. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018)
10. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection pipeline and challenges for epic-kitchens-100. IJCV (2021)
11. Dong, Q., Gong, S., Zhu, X.: Imbalanced deep learning by minority class incremental rectification. IEEE TPAMI **41**(6), 1367–1381 (2018)
12. Escorcia, V., Carlos Niebles, J.: Spatio-temporal human-object interactions for action recognition in videos. In: ICCVW (June 2013)
13. Escorcia, V., Soldan, M., Sivic, J., Ghanem, B., Russell, B.C.: Temporal localization of moments in video collections with natural language. CoRR **abs/1907.12763** (2019), http://arxiv.org/abs/1907.12763
14. Falcon, W.: Pytorch lightning. https://github.com/PytorchLightning/pytorch-lightning (2019)
15. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)
16. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: CVPR (2018)
17. Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., Bojanowski, P.: Self-supervised pretraining of visual features in the wild. CoRR (2021), https://arxiv.org/abs/2103.01988
18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS (2020)
19. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (Sep 2020). https://doi.org/10.1038/s41586-020-2649-2, https://doi.org/10.1038/s41586-020-2649-2

20. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
22. Hou, Z., Peng, X., Qiao, Y., Tao, D.: Visual compositional learning for human-object interaction detection. In: ECCV (2020)
23. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: CVPR. pp. 5375–5384 (2016)
24. Huang, Y., Cai, M., Li, Z., Lu, F., Sato, Y.: Mutual context network for jointly estimating egocentric gaze and action. IEEE TIP **29**, 7795–7806 (2020)
25. Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J.: Exploring balanced feature spaces for representation learning. In: ICLR (2021)
26. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: ICLR (2020)
27. Kazakos, E., Huh, J., Nagrani, A., Zisserman, A., Damen, D.: With a little help from my temporal context: Multimodal egocentric action recognition. In: BMVC (2021)
28. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (2019)
29. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Slow-fast auditory streams for audio recognition. In: ICASSP (2021)
30. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: Ego-exo: Transferring visual representations from third-person to first-person videos. In: CVPR (2021)
31. Li, Y., Liu, M., Rehg, J.: In the eye of the beholder: Gaze and actions in first person video. IEEE TPAMI (2021)
32. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: ECCV (2018)
33. Li, Y.L., Liu, X., Wu, X., Li, Y., Lu, C.: Hoi analysis: Integrating and decomposing human-object interaction. In: NeurIPS (2020)
34. Lin, J., Gan, C., Han, S.: Temporal shift module for efficient video understanding. In: ICCV (2019)
35. Liu, M., Tang, S., Li, Y., Rehg, J.M.: Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In: ECCV (2020)
36. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: ICLR (2021)
37. Narasimhaswamy, S., Nguyen, T., Hoai, M.: Detecting hands and recognizing physical contact in the wild. In: NeurIPS (2020)
38. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) NIPS, pp. 8024–8035 (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
39. Purushwalkam, S., Gupta, A.: Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. In: NeurIPS (2020)
40. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.: Learning human-object interactions by graph parsing neural networks. In: ECCV (2018)

41. Reed, C.J., Yue, X., Nrusimha, A., Ebrahimi, S., Vijaykumar, V., Mao, R., Li, B., Zhang, S., Guillory, D., Metzger, S., Keutzer, K., Darrell, T.: Self-supervised pretraining improves self-supervised pretraining. arXiv:2103.12718 (2021)
42. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
43. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: CVPR (2020)
44. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and observer: Joint modeling of first and third-person videos. In: CVPR (2018)
45. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
46. Sudhakaran, S., Escalera, S., Lanz, O.: Lsta: Long short-term attention for egocentric action recognition. In: CVPR (2019)
47. Sudhakaran, S., Lanz, O.: Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In: BMVC (2018)
48. Sun, C., Nagrani, A., Tian, Y., Schmid, C.: Composable augmentation encoding for video representation learning. In: ICCV (2021)
49. Umesh, P.: Image processing in python. CSI Communications **23** (2012)
50. Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., Murphy, K.: Tracking emerges by colorizing videos. In: ECCV (2018)
51. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE TPAMI **41**(11), 2740–2755 (2018)
52. Wang, X., Wu, Y., Zhu, L., Yang, Y.: Symbiotic attention with privileged information for egocentric action recognition. In: AAAI (2020)
53. Wang, X., Zhu, L., Wang, H., Yang, Y.: Interactive prototype learning for egocentric action recognition. In: ICCV (2021)
54. Wang, X., Gupta, A.: Videos as space-time region graphs. In: ECCV (2018)
55. Wang, X., He, K., Gupta, A.: Transitive invariance for self-supervised visual representation learning. In: ICCV (2017)
56. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: CVPR (2019)
57. Yan, R., Xie, L., Shu, X., Tang, J.: Interactive fusion of multi-level features for compositional activity recognition. arXiv:2012.05689 (2020)
58. Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. NeurIPS **33**, 19290–19301 (2020)
59. Zhang, X., Wu, Z., Weng, Z., Fu, H., Chen, J., Jiang, Y.G., Davis, L.S.: Videolt: Large-scale long-tailed video recognition. In: ICCV. pp. 7960–7969 (2021)
60. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV (2018)