Supplemental Material: 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

We complement the manuscript with the following items:

– Results in the test set of EPIC-KITCHENS-100 without public labels, refer to Section 1.
– Additional results of two model ensemble ablations, refer to Section 2.
– Additional qualitative results, refer to Section 3.
– A pragmatic take to understand quantitative differences in Epic-Kitchens 100 [3], refer to Section 4.
– A list of Frequently Asked Questions (FAQ), refer to Section 5.

1 Results in EPIC-KITCHENS-100 test set

To validate the generalization of the insights presented in the main paper, we resort to the test set of EPIC-KITCHENS-100 with sequestered labels. Concretely, we submitted the predictions of a given model onto the evaluation server. Table 1 reports the results from the evaluation server, with the metrics established by the dataset creators [3].

Overall the observations drawn from the validation can be seen in the test set with sequestered data. We repeat some of the main observations here: **(1)** The performance of CNN action models, TSM, is similar to the recently introduced X-ViT without heavy test-time data augmentation. **(2)** Importantly, our OIC further improves X-ViT by 1.6%, 5.4%, and 2.6% on overall verb, noun, and action accuracy, respectively. It is also seen that the improvement is mainly achieved in noun recognition as expected. Thus, it indicates that the attention mechanism of transformer is limited in extracting active objects in manipulation and interaction with human hands from cluttered background and scene. 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)** The recent X-ViT model significantly 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 1.2% on overall actions to this model.

Modality	Method	TTDA	Overall Top-1 Verb	Overall Top-1 Noun	Overall Top-1 Action	Overall Top-5 Verb	Overall Top-5 Noun	Overall Top-5 Action	Unseen participants Top-1 Verb	Unseen participants Top-1 Noun	Unseen participants Top-1 Action	Tail classes Top-1 Verb	Tail classes Top-1 Noun	Tail classes Top-1 Action
V	X-ViT† [1]	✗	58.5	46.9	34.5	85.2	71.4	53.0	50.4	37.7	25.7	28.2	21.9	14.0
V	+**OIC**	✗	**60.1**	**52.3**	**37.1**	**87.7**	**77.7**	**59.0**	**53.0**	**44.9**	**28.8**	**30.4**	**25.2**	**16.1**
V	X-ViT [1]†	✓	64.3	52.9	40.3	88.1	77.4	60.6	58.2	45.7	32.0	31.2	26.3	17.7
V	+**OIC**	✓	**64.7**	**55.0**	**41.5**	**89.2**	**79.5**	**62.7**	**58.6**	**48.0**	**33.0**	**31.4**	**27.6**	**18.1**
V	TSM† [5]	✗	58.5	45.9	32.8	86.4	72.3	54.0	53.1	41.1	27.1	27.9	21.5	13.8
F	TSM† [5]	✗	60.9	33.8	26.2	86.1	57.6	43.5	55.5	27.5	20.1	26.0	6.4	7.5

Table 1. Results on the test set of EPIC-KITCHENS-100 [3] with models only trained in the training set. Modality: V=Visual; F=Optical flow. †: Results computed with the model weights released by the authors of [3,1]. TTDA: Test-time data-augmentation (*e.g.*, multi-crop). Underlined numbers correspond to the best results across the board. Bold numbers highlight the best between a proxy model and proxy + ours. All in all, our model complements X-ViT and yields state-of-the-art results.

2 Additional ablation of two models' ensemble

Our full video prediction network corresponds to the fusion of our OIC Net and one standard video classification network, *cf.* main paper Sec. 3.2 and Fig 4. The intuitions behind the proposed fusion are: (1) it allows to validate the relevance of handled objects and learn a specialized neural network representation for them. (2) It is a simple way to integrate the information captured by our OIC Net without the need for model re-training. (3) It combines the complementary information discovered by our OIC Net relating to objects while the video net provides the scene context. For a fair comparison, we compare the results of our fused model w.r.t the proxy video network used in our fusion and the ensemble of two networks of a given proxy model.

Implementation details. We used the standard TSM video architecture [5] with the pre-trained weights from [3] as proxy video network. For the two model ensemble baseline, we train an additional TSM model using the code and training recipe given by [3]. We followed the training pipeline presented in the main paper for our fused model. Table 2 reports the experimental results in the validation set of EPIC-KITCHENS-100 with the metrics established by the dataset creators [3]. We have the following observations: (1) TSM ensemble yield an absolute performance improvement of +0.7%, +1.0% and +0.6% on overall verb, noun and action respectively. Meanwhile, our model fusion with OIC yields significantly better absolute improvement for noun and action (+5.0% and +3.4%) and slightly better absolute improvement for verb (+1.0%). (2) We also observe that our OIC improves the accuracy scores on unseen participants (for noun and action) and tail classes (for verb, noun, and action). As expected, the most significant performance improvement is from noun accuracy.

In summary, compared to a strong model ensemble baseline, we have validated the significance of the results from our fused model, with OIC.

Method	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
TSM† [5]	63.0	47.3	35.6	88.9	74.2	57.2	54.6	37.1	26.3	35.9	26.7	18.1
TSM ensemble	63.7	48.3	36.2	89.7	75.3	58.6	**55.3**	38.5	26.8	34.7	26.3	18.0
TSM + **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**

Table 2. Results on the validation set of EPIC-KITCHENS-100 [3]. †: Results computed with the model weights released by the authors of [3]. Bold numbers highlight the best between a proxy model (TSM), two model ensembles of the proxy model (TSM ensemble), and proxy + ours (TSM + OIC). All in all, our OIC improves further the proxy model validating the relevance of our handled objects representation.

3 Qualitative results

Fig. 1 depicts additional qualitative results to those shown in Fig. 5 in the *cf.* main manuscript. Each row depicts five evenly spaced frames from a particular example with its associated object regions, the ground truth labels, predictions made by TSM, and predictions made by TSM+OIC (ours). As before, correctly predicted verbs and nouns are highlighted as green, while incorrectly predicted as red. These additional results further reinforce the notion that the additional context provided by the OIC module guides prediction correctly. Fig. 1 (*a* - *f*), depicts examples where the noun of interest is either very small (*a*, *b*, *e* and *f*), out of view during a large portion of the clip (*b* and *d*) or not well illuminated (*c*). Nevertheless, including the OIC module allows this information to be correctly captured and preserved, while is lost or ignored with pure TSM. Fig. 1 , (*g*) depicts an example where both TSM and ours make a wrong prediction, in the case of TSM the verb is correctly identified, however, the noun predicted is very different from the ground truth. In this same example, our method correctly predicts the noun, while incorrectly predicting the verb; however, one could argue that the verb wrongly predicted by our method could apply to the example, while the same cannot be said by the wrongly predicted noun by TSM. Similar arguments as the ones just discussed for Fig. 1 (*g*) would apply to *h*, where additionally TSM predictions of both verb and noun are incorrect. Finally, Fig. 1 (*j* - *k*), shows two fail examples for both methods, where although our method makes use of object boxes largely associated with the target noun, the proposed method still fails.

Fig. 2 shows examples of patches representing objects used during SOS pre-training stage, *i.e.* to train our proposed SwAV-S. The figure also displays the associated labels from the video for reference. Still, it is worth noting that our approach does not exploit the given labels during the SOS pre-training. There, it can be appreciated the *real* spatiotemporal transformation that objects experience. These, represent the sets of objects presented to SwAV-S for cluster assignment and prediction. Fig.2 left column illustrates cases where the action process can easily be seen as the primary source of variation (*e.g.*, the first row exhibits large deformations of a flattened box). In contrast, the right column

Fig. 1. Qualitative results. The green bounding boxes represent automatically detected object regions being manipulated. Rows (a - f) represent examples that are corrected by our method, rows (g - i) examples that are partially correct, while rows (j - k) depict failure cases for both methods.

illustrates more challenging examples that can capture the semantics of the underlying action. The examples presented in Fig.2 left, also elucidate why our approach still works without tracking or strictly enforcing object correspondences across frames.

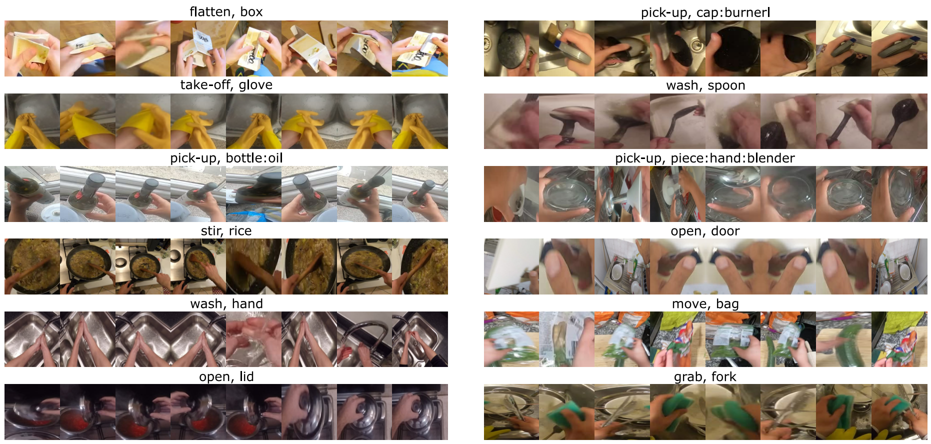

Fig. 2. Qualitative results. SwAV-S input objects and their related *real* spatiotemporal transformations.

4 Performance improvements perspective

Recently, Picard studied the influence of random seeds in modern deep learning architectures in computer vision [6]. From their study, the relevance and importance of setting a confidence interval to gauge meaningful improvements in the performance of deep neural networks are brought to the forefront. However, it is difficult to define rigorous statistical confidence intervals in practice, as DNN often takes a very long time to train. For example, a complete training cycle of our approach (SSL pre-training and target task fine tuning) could take a handful of days with a sub-optimal implementation and/or system architecture. Thus, we resort to a pragmatic approach to define a confidence interval. During the initial stage of this project, we ran our simplest baseline five times with the same hyperparameter configuration, but with different initialization seeds. By contrasting the differences in performance among the multiple performance metrics (KPIs) across multiple runs, we observe that KPIs varied in a range $\pm 0.25\%$. Therefore, we deem relevant results those bigger than 0.5%.

5 FAQ

We genuinely appreciate the feedback from our reviewers (CVPR & ECCV). This section includes some relevant questions/concerns/doubts raised during the peer review process that curious readers may ask themselves.

1. **Why did we term our approach "Self-supervised learning Over Sets (SOS)"?**
   Answer. SOS relies on the self-supervisory signal from a collection of independent and automatically detected object proposals, representing the "set" of

objects and their temporal deformations. This distinctly differs from SwAV [2] and standard image-based SSL algorithms, where the synthetic transformations (augmentations) operate over a single image.

2. **Principles behind SOS. What is intuition? Why concurrence?**
Answer. As mentioned in the Introduction and Section 3.1 (*cf.* main submission), SOS leverages the notion of actions as natural transformations. For example, an action like "cutting onion" implies non-rigid deformations of "an onion" difficult to replicate by the current standard data augmentation techniques. Modeling "actions" as transformations equates to treating them as an implicit conditioning signal enriching the visual appearance of handled objects. It is worth noting that the object proposals given to SOS may be visually different and involve a single object category (*cf.* Fig. 2 row 1 of the left column - "flatten box") or multiple objects (*cf.* Fig. 2 row 4 of the left column - "stir rice") object categories. As such, SOS exploits two self-supervisory signals: **(1.)** temporal continuity, and **(2.)** concurrence. Further, concurrence relations are statistically/implicitly learned, by the network as it is forced to map proposals that co-occur often (e.g.,"pan" and "spoon"). Here, we favor the term"temporal continuity" over "temporal consistency" as natural transformations are not necessarily rigid.
3. **SwAV choice.**
Clarification. Preliminary experiments showed that SwAV offered the best performance, e.g. outperforming SimCLR by +2.7/3.0 in noun/action. However, since the method evolved from that preliminary experiment, we extended the ablation in Table 4 (*cf.* main submission). SwAV still offers a better initialization than SimCLR, +0.7/+1.1 for noun/action. MoCoV3 offers +0.0/+0.3 in noun/action vs SwAV. SwAV pre-trained weights are empirically shown to be a reasonable choice.
4. **Marginal improvements**.
Clarification. We are sorry that you think so. Kindly note that by aiding three strong action recognition models with our OIC the results for top-1 accuracy improve as follows +5.0/+3.4 noun/action (TSM), +3.5/+2.4 (SlowFast) and +5.5/+3.2 (XViT), *i.e.*the improvement is not *marginal*. It is worth mentioning that we also compared our approach w.r.t action modeling active objects [8]. Also kindly consider reading the Section 4 to gauge improvements in EPIC-KITCHENS-100 properly.
5. **Is the experimental comparison unfair? The model's improvements are expected as the final results correspond to a fused model**.
Answer. We strive for fair experimentation and comparison. We resort to a simple yet effective model fusion to showcase the relevance of **(1)** paying more attention to handled objects, and **(2)** learning an appropriate representation of handled objects. As such, the performance gains w.r.t three proxy action recognition models [1,4,5] (five if someone considers TSM-RGB+Flow and XViT-TTDA as additional flavors) without OIC, reflect the benefit of appropriate modeling of handled objects.
We have included an ablation showcasing the relevance of our fused model against a strong two-model ensemble baseline in Section 2.

6. **Did you disregard the pre-trained weights of the projection and prototype subnetworks of SwAV during the Stage II (SOS pre-training)?**
   Answer. As mentioned in Section 4 (*cf.* main submission), we only re-used the pre-trained weights corresponding to the CNN backbone (*i.e.*the encoder). The rationales behind our choice are: (1) our problem setup involves domain-adaptation (from natural to egocentric images), and (2) the fact that SwAV is an online clustering approach [2]. Thus, we did not consider reusing the prototypes specialized for reasoning about natural images.
7. **Is not the multi-stage training cumbersome?**
   Answer. Stage-I corresponds to using standard readily available pre-training weights (*i.e.*torch.load). Subsequent stages correspond to on-domain training. Most work either train from an ImageNet pre-trained model (1 less stage) or use an on-domain pre-train+finetune approach (e.g. ImageNet → Kinetics → EpicKitchens/UCF101/SSV2), resulting in the same number of stages as ours. Hence, our method is very similar in training complexity to a standard two-stage approach. It is relevant to highlight that our approach is closer to representation learning than other areas of research in neural networks (*e.g.*, downstream specialization, or architectural design).
8. **Number of objects from HOI.**
   Clarification. It is worth noting that the hand-object detector [7] predicts more than one object proposal for each hand. [7] employs a heuristic based on proximity and confidence for visualization purposes. In our work, we use all the object proposals gathered after the filtering steps described in Section 4 (*cf.* main submission).
9. **The paper only validates a single target task fine-tuning step.**
   Clarification. We believe that our work has applicability to general action recognition tasks interacting with objects, *e.g.*, temporal action localization, instructional videos, *etc.*, since we a) obtain a strong representation of handled objects and b) can finetune with video-level labels (no bounding box labels are required). However, reusing off-the-shelf models is favored per task due to computing limitations. Thus, we believe our OIC is handy.
10. **Computational complexity and runtime.**
    Clarification. We benchmarked our approach in a 1080Ti GPU. TSM and XViT run in 12 and 66 msec. respectively, while the OIC module runs in 10.7 msec. The OIC complexity is 26 GFLOPs while XViT is much larger, 285 GFLOPs. Thus, our OIC is comparatively efficient. It is worth noting that we computed our run-time and computational complexity based on a naive implementation feeding cropped handled object patches to the CNN instead of employing ROIPooling or ROIAlign layers.
11. **Are the results marginal in regards to Test-Time Data Augmentation (TTDA)?**
    Answer. TTDA adds much more computational complexity (*cf.* FAQ-Q10). It is worth noting that: (1) our approach is orthogonal, and offers improved performance. (2) We do **not** aim to substitute TTDA. (3) A combination

of TTDA and our approach might offer further gain, but we believe it is beyond our work's scope.

References

1. Bulat, A., Perez-Rua, J.M., Sudhakaran, S., Martinez, B., Tzimiropoulos, G.: Space-time mixing attention for video transformer. In: NeurIPS (2021)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
3. 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)
4. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)
5. Lin, J., Gan, C., Han, S.: Temporal shift module for efficient video understanding. In: ICCV (2019)
6. Picard, D.: Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision (2021)
7. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: CVPR (2020)
8. Wang, X., Zhu, L., Wang, H., Yang, Y.: Interactive prototype learning for egocentric action recognition. In: ICCV (2021)