Supplementary Material for Ego-Only: Egocentric Action Detection without Exocentric Transferring

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

In this supplementary material, we provide more dataset statistics, implementation details, ablations, and analyses.

1. Dataset Statistics

Ego4D [7] offers 3,670 hours of daily life egocentric videos from hundreds of scenarios, providing massive-scale data for self-supervised pretraining. The Ego4D Moments Queries (MQ) task in the Episodic Memory benchmark contains 110 moments classes, 326.4 hours of videos (194.9h in train, 68.5h in val, 62.9h in test), 2522 clips (1486 in train, 521 in val, 481 in test), and 22.2K annotated temporal action segments (13.6K in train, 4.3K in val, 4.3K in test).

EPIC-Kitchens-100 [4] offers 100 hours (74.7h in train, 13.2h in val, 12.1h in test) of egocentric videos from 700 sessions (495 in train, 138 in val, 67 in test) in 45 kitchens. The Action Detection challenge contains 97 verb classes (97 in train, 78 in val, 84 in test), 300 noun classes (289 in train, 211 in val, 207 in test), and 90.0K temporal action segments (67.2K in train, 9.7K in val, 13.1K in test).

Charades-Ego [11] offers 8K videos (3K in ego train, 3K in exo train, 846 in ego test) of daily indoor activities. The videos are recorded from both third and first person with temporal segments annotated (33K in ego train, 34K in exo train, 9K in ego test) over 157 classes.

2. Implementation Details

MAE pretraining. As discussed in Section 3.1 of the main paper, we follow the technical details in video MAE [6] unless noted otherwise. However, as egocentric datasets contain long videos with hundreds or thousands of hours, in this paper, we define one epoch as 245,760 clips sampled from data, so that the compute budget is comparable to one Kinetics-400 [9] epoch. With this definition, we pretrain egocentric MAE for 800/1600 epochs, batch size 256, without repeated sampling for simplicity, learning rate 8e-4, by default. We sample clips of 16 frames with a temporal span of 2 seconds, equivalent to a sampling rate of 4 in 30-fps videos.

Finetuning. We finetune for 20 epochs with 2-epoch warm-up, batch size 128, RandAugment [3], stochastic depth [8] 0.2, dropout [12] 0.5, label smoothing 0.0001 for BCE, no mixup [16] or cutmix [14] as they are not common for segmentation. We use SGD with learning rate 4.0 weight decay 0.0 on Ego4D, while we use AdamW [10] with learning rate 8e-4 weight decay 0.05 on EPIC-Kitchens-100. For finetuning on EPIC-Kitchens-100, we concatenate all verb and noun classes so that we finetune only once.

Action detection. As discussed in Section 3.1 of the main paper, we follow the details of ActionFormer [15] for EPIC-Kitchens-100 unless noted otherwise. Our Ego4D features are extracted at stride 8 which equals the transformer output stride, with frame sampling rate 4 and temporal patch stride 2. The sliding windows use stride 8 as well. We train for 10 epochs with 8-epoch warm-up, learning rate 2e-4. EPIC-Kitchens-100 features use stride 16 [15] for fair comparison. We train for 20 epochs with 16-epoch warm-up, learning rate 2e-4. We report an average of 3 runs.

Action recognition. We sample clips of 32 frames [13] with a temporal span of 3.2 seconds, equivalent to a sampling rate of 3 in 30-fps videos. And due to the extra memory constraint, we reduce the batch size to 64. On Charades-Ego without exocentric data, we train 10 epochs with 1-epoch warm up, SGD optimizer, learning rate 0.8, and no weight decay. On Charades-Ego with exocentric data, we instead use AdamW optimizer, learning rate 2.4e-4, and weight decay 0.05. On EPIC-Kitchens-100, we train 20 epochs with 2-epoch warm up, AdamW optimizer, learning rate 2.4e-4, and weight decay 0.05.

3. Ablation on Concatenated Features

In Figure 1, we present the ablation of concatenating features from the last few (2, 3, 6, or 12) transformer blocks, instead of our default choice of the last block only. This is inspired by the linear protocol in DINO[2] that was aimed to improve results with frozen self-supervised learning features (in our case frozen MAE features) but we ablate this choice for all models, with and without finetuning. How-

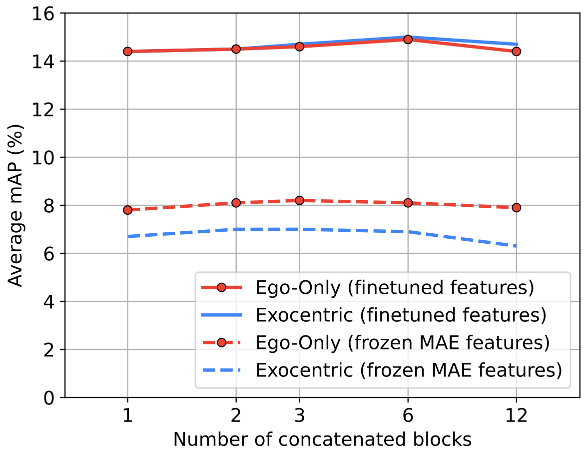

Figure 1. Ego4D Moments Queries results with concatenated features from the last few (2, 3, 6, 12) transformer blocks (12 blocks in total for the ViT-B [5] architecture), instead of our default choice of the last block only. The detection results are almost not affected in any of the four models studied. This stable gap between finetuned features and frozen MAE features verifies the necessity of the egocentric finetuning stage in Ego-Only.

method	MAE	exo	rebalancing technique	mAP
exo-FT	K400	K400	resampling	16.2
exo-FT	K400	K400	per-class reweighting	14.4
exo-FT	K400	K400	per-instance reweighting	16.2
Ego-Only	Ego4D	-	resampling	16.3
Ego-Only	Ego4D	-	per-class reweighting	14.4
Ego-Only	Ego4D	-	per-instance reweighting	16.3

Table 1. Varying rebalancing techniques. Ego-Only matches exocentric transferring regardless of rebalancing techniques.

ever, we see a marginal gain for frozen MAE features, which confirms the necessity of the egocentric finetuning stage in Ego-Only.

4. Ablation on Rebalancing Techniques.

As discussed in Section 3.2 of the main paper, we are currently mitigating the imbalance challenges by simply reweighting the loss according to the number of positive frames in each action instance. Beyond this current technique, we also study a simple action resampling option as a natural alternative. Specifically, instead of uniformly sampling all the clips within the train data, we sample only the center 2 seconds of each action regardless of the action length, similar to an action classification task. As shown in Table 1, this resampling option performs the same as the default reweighting with and without exocentric transferring. We also study a per-class reweighting method that ignores action length imbalance within a class and find that it performs worse than the other two rebalancing methods. In all these cases, our Ego-Only method matches Kinetics transferring, without any exocentric data or label, and *regardless* of the rebalancing techniques employed. We consider further exploration of better rebalancing methods as an open research problem and leave it to future work beyond the scope of this paper.

5. Error Analyses

False positive analysis. In Figure 2, we analyze false positive errors on EPIC-Kitchens-100 [4] with ViT-L [5] models using the DETAD [1] error diagnosing tool. The models are trained with per-class reweighting. We notice that Ego-Only reduces false positive errors on backgrounds, compared with exocentric pretraining baselines, probably because Kinetics [9] contains mostly trimmed videos with foreground actions only.

Sensitivity analysis. In Figure 3, we analyze the model sensitivity according to DETAD characteristics [1] on EPIC-Kitchens-100 [4] with ViT-L [5] models. The models are trained with per-class reweighting. We observe that our Ego-Only improves significantly when there are multiple verb instances of the same category in a video.

References

[1] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *ECCV*, 2018. 2, 3, 4

[2] 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. 1

[3] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020. 1

[4] 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] 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

[6] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 1

[7] 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

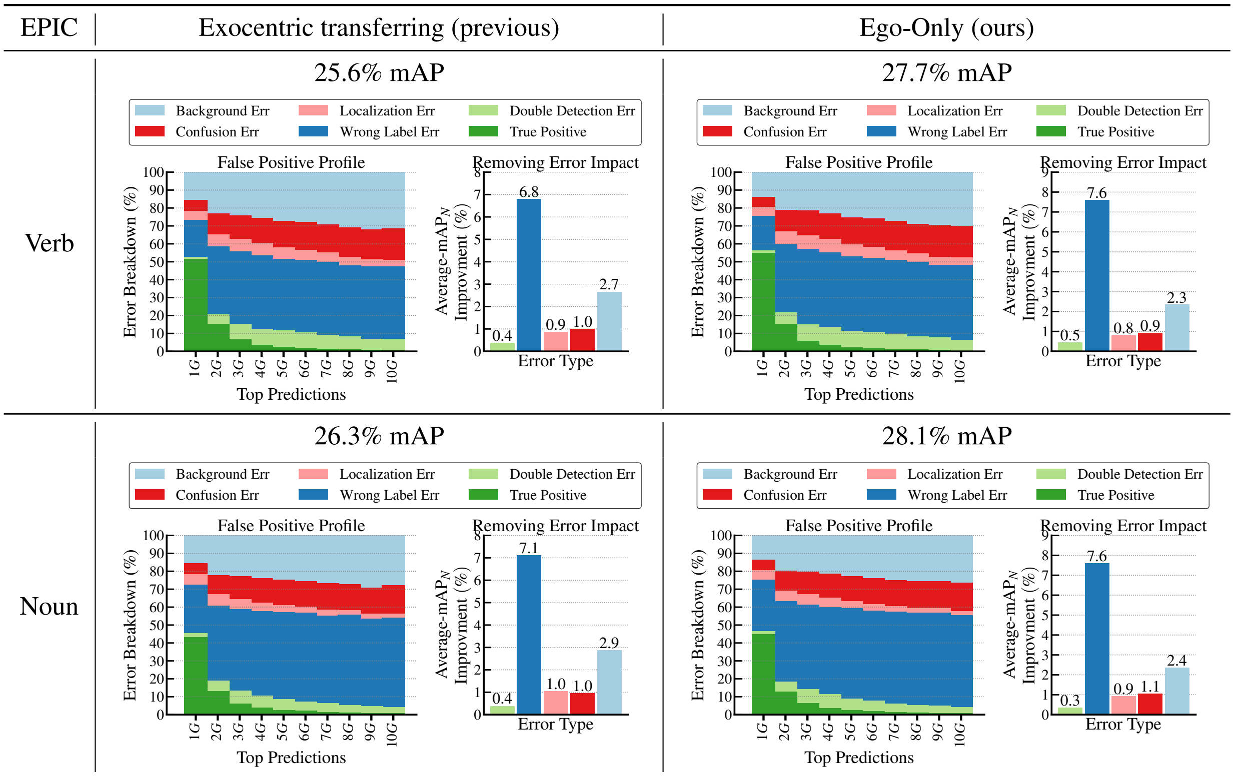

Figure 2. False positive analysis on EPIC-Kitchens-100 [4] with DETAD [1]. The error types are determined by the tIoU between ground-truth and predicted segments, as well as the correctness of the predicted labels. Background error: tIoU $<$ 1e-5; confusion error: 1e-5 $<$ tIoU $< \alpha$ and label is wrong; localization error: label is correct but 1e-5 $<$ tIoU $< \alpha$; wrong label error: tIoU $>= \alpha$ but label is wrong, where α refers to the tIoU thresholds $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. 'G' refers to the number of ground-truth instances. According to the error breakdown, although the large-scale exocentric pretraining helps reducing wrong label errors, our Ego-Only predicts more true positives correctly and reduces background errors, probably because Kinetics [9] contains mostly trimmed videos with foreground actions only.

[8] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 1

[9] 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, 3

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[11] 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

[12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014. 1

[13] 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. 1

[14] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 1

[15] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *ECCV*, 2022. 1

[16] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1

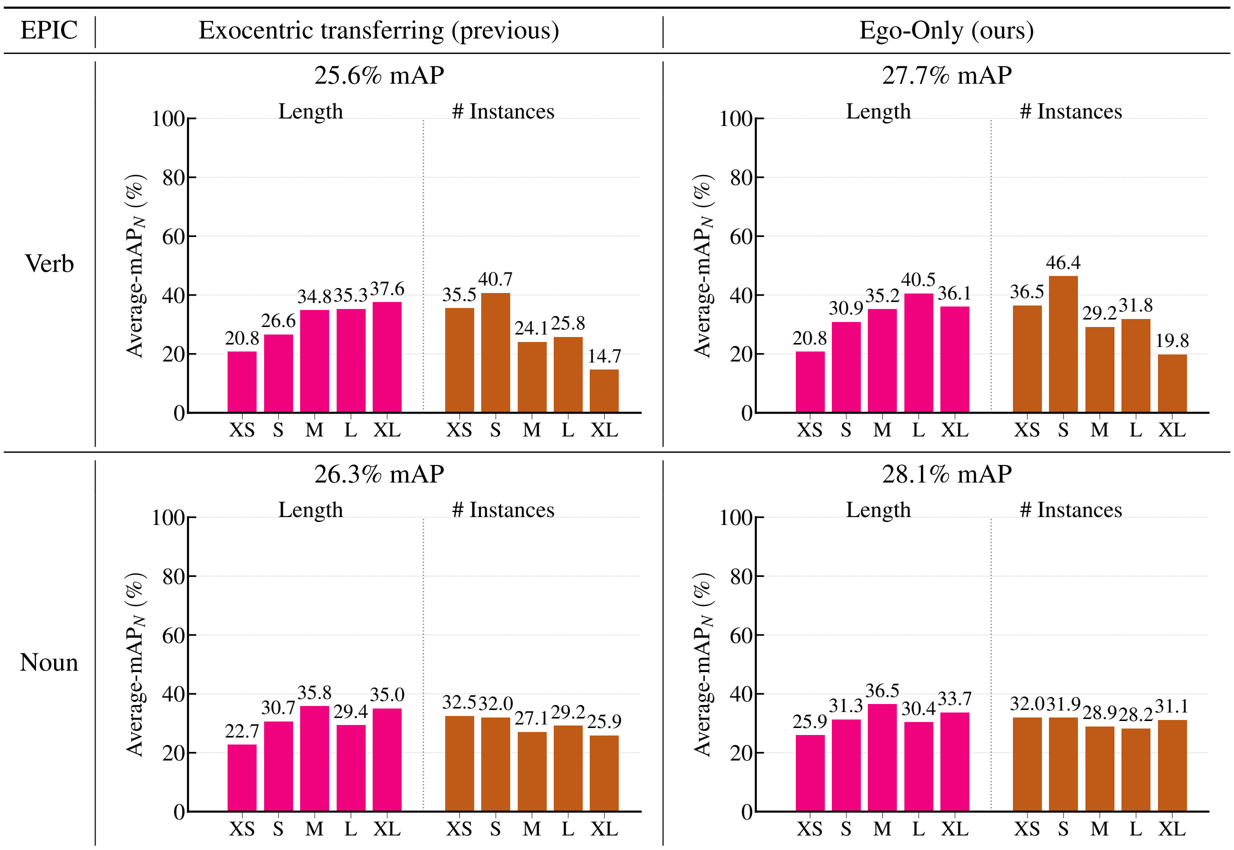

Figure 3. Sensitivity analysis on EPIC-Kitchens-100 [4] with DETAD [1]. Ground-truth segments are divided into 5 equal buckets according to their characteristic [1] percentiles. Then, average mAP_N [1] metrics are computed for each characteristic bucket. The 'length' characteristic measures the length of the ground-truth action segment in seconds. The '# instances' characteristic measures the number of action instances belonging to the same category as the ground-truth segment in the same video. According to the average mAP_N in each bucket, we observe that our Ego-Only improves significantly when there are multiple verb instances of the same category in a video.