Leveraging Next-Active Objects for Short-Term Anticipation in Egocentric Videos : Supplementary Material

Sanket Thakur[1,4], Cigdem Beyan[2], Pietro Morerio[1], Vittorio Murino[3, 1], and Alessio Del Bue[1]

[1]Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT)
[2]Department of Management, Information and Production Engineering, University of Bergamo, Dalmine, Italy
[3]Department of Computer Science, University of Verona, Italy
[4]Department of Electrical, Electronics and Telecommunication Engineering and Naval Architecture (DITEN), University of Genoa, Italy

This supplementary material presents the qualitative analysis of our model, NAOGAT on Ego4D [15] and EpicKitchen-100 [4] dataset. We provide a video depicting the performance of our model when progressed over the allowed the observed segment of a video clip, which is discussed in detail in Sec. 1. In addition, we also provide some visualization for next-active-object (NAO) annotation on EpicKitchen-100 [4], depicting its location and the class label in the last observed frame for a given video clip. We also describe the annotation pipeline followed to curate the ground-truth data for next-active-object prediction for the Short-Term Anticipation task in Sec. 2.

1. Video

We provide additional detail on performance of our model, NAOGAT, when compared with the object detections provided by the object detector pre-trained on Ego4D [15]. We notice a significant improvement in refining the object detections and also identifying objects which are not detected by the object detector to anticipate the location of NAO. The video entails the performance of NAOGAT auto-regressively when fed with a sequential progressive video clip. It can be noticed that as the video progresses, the model further refines the predictions based on past observations and predicts the next-active-object bounding box and its class label, along with future action and time to contact (TTC) with the object. The video also provide a visualization on future frames which are not observed by the model describing the time taken to contact with the next-active-object.

2. EpicKitchen-100 NAO dataset curation

The Short-Term Anticipation (STA) task involves predicting the location (bounding box, $\hat{b}$) and class label, $\hat{n}$ of the next-active-object, as well as the future action $\hat{v}$ and the time to contact (δ) with the NAO, for a given video clip. It is important to note that the NAO must be present and visible in the last observed frame for the task to be valid. Currently, only Ego4D [15] dataset provides the precise annotation for studying the problem.

The EpicKitchen-100 dataset [4] offers valuable ground-truth data for the action anticipation [12, 13] task. The dataset includes information on future actions such as ”peeling an onion,” future verbs like ”peel,” and associated noun labels of the object involved in the action, such as ”onion.” This makes the dataset an excellent resource for studying and evaluating models designed to predict future actions. We consider the noun label as the NAO class label for a given clip. However, it lacks annotations for the location of NAO in the last observed frame. For this purpose, we aimed to curate our own annotation for NAO estimation following the pipeline described in Fig. 1.

To curate ground-truth data for the next-active-object prediction for the Short-Term Anticipation task, we first extract the last observed frame from a given clip. Next, we use a pre-trained object detector [35] on the EK-55 dataset [5] to obtain raw object detections for the frame. We then verify if the ground-truth NAO class label is identified in the raw detections. If a match is found, the corresponding bounding box for that detection is used as the ground-truth annotation for the NAO bounding box ($\hat{b}$). However, if the object detector fails to identify any object with the ground-truth NAO

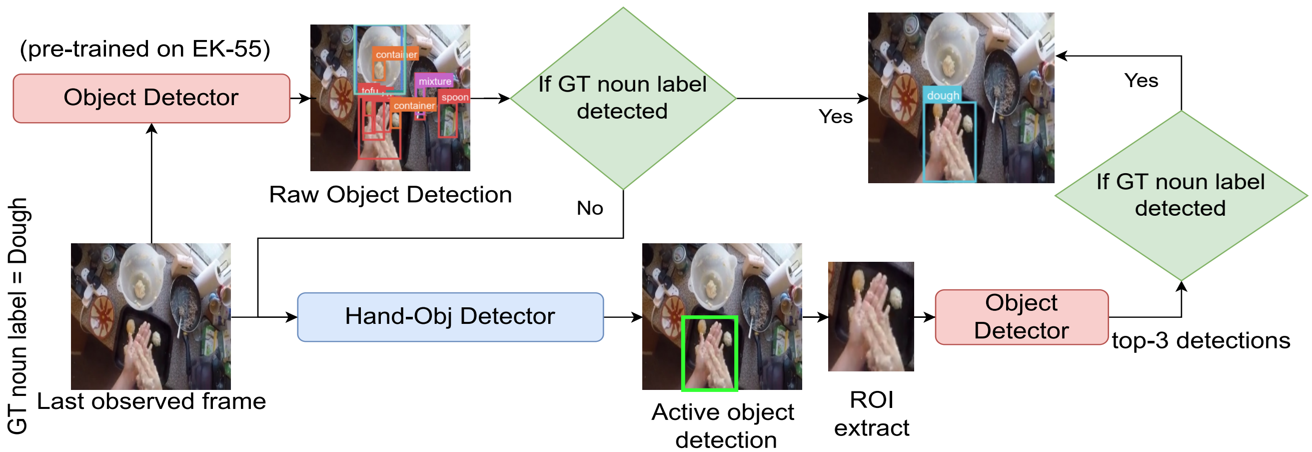

Figure 1. Annotations pipeline for extracting next-active-object ground-truth labels for EpicKitchen-100 [4] dataset.

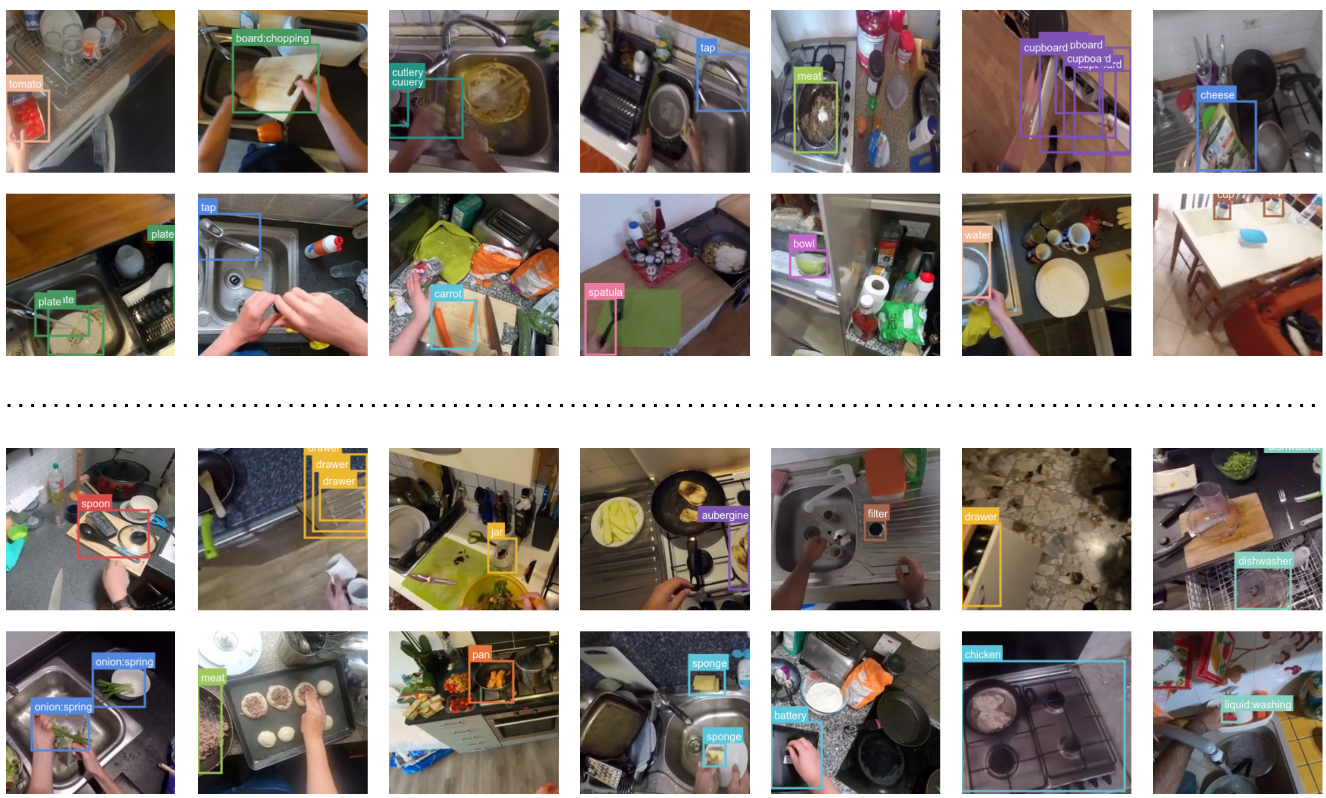

Figure 2. NAO annotations for EK-100 as curated from the pipeline described in Fig. 1. The frames corresponds to the last observed frame for a given clip and the detection represents the next-active-object information in terms of NAO location and its class label.

label, we use a Hand-Object detector [?] to obtain bounding boxes for the active object [32]. This is because the hand-object detector has been shown to be state-of-the-art in identifying hand-object detection and has been used in the literature [24, 37]. In the event that the Hand-Object detector identifies an active object, we extract the Region of Interest (ROI) for the corresponding detection from the input frame. This ROI is then fed into the object detector [35] used earlier, and we take the top-3 predictions from the detector. These predictions are once again verified against the Ground-Truth NAO class label to check if they contain the NAO label. If one of the predictions satisfies the criteria, the location of the active object is used as the ground-truth annotation for the NAO location. This pipeline is used to only curate information regarding the location of NAO and not the class label of NAO for a given clip. The class label for NAO is used from the annotations provided with EK-100 for action anticipation. The final annotations for the dataset are shown in Fig. 2.

3. Limitations of our model for EpicKitchen-100 dataset

It is important to note that EpicKitchen-100 was not curated in alignment with the definition of STA. Specifically, the dataset does not provide annotations for next-active-object, and it is not mandatory for NAO to be present in the allowed last frame observed by the model. As discussed in the main paper, our dataset curation method (described in Sec. 2) could not annotate the ground-truth data for the next-active-object in *22%* of the "Test Set" of the Validation split, as there were no detected objects in those clips. Moreover, the EK-100 dataset suffers from a dataset bias, as there are *300* class labels for objects, and similar-looking objects are often classified differently, as shown in Fig. 3. This further confuses the model's identification of objects and impedes its ability to anticipate future actions.

References

[1] Anna M Borghi. Object concepts and action. *Grounding cognition: The role of perception and action in memory, language, and thinking*, pages 8–34, 2005.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing.

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *International Journal of Computer Vision*, 2021.

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[7] Eadom Dessalene, Chinmaya Devaraj, Michael Maynord, Cornelia Fermuller, and Yiannis Aloimonos. Forecasting action through contact representations from first person video. *IEEE TPAMI*, pages 1–1, 2021.

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

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[11] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017.

[12] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *International Conference on Computer Vision*, 2019.

[13] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *ICCV*, 2021.

[14] Ross Girshick. Fast r-cnn. In *IEEE ICCV*, pages 1440–1448, 2015.

[15] Kristen Grauman, Andrew Westbury, and Eugene et al. Byrne. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] Julia Hertel, Sukran Karaosmanoglu, Susanne Schmidt, Julia Bräker, Martin Semmann, and Frank Steinicke. A taxonomy of interaction techniques for immersive augmented reality based on an iterative literature review. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 431–440, 2021.

[18] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3148–3159, June 2022.

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[20] Jingjing Jiang, Zhixiong Nan, Hui Chen, Shitao Chen, and Nanning Zheng. Predicting short-term next-active-object through visual attention and hand position. *Neurocomputing*, 433:212–222, 2021.

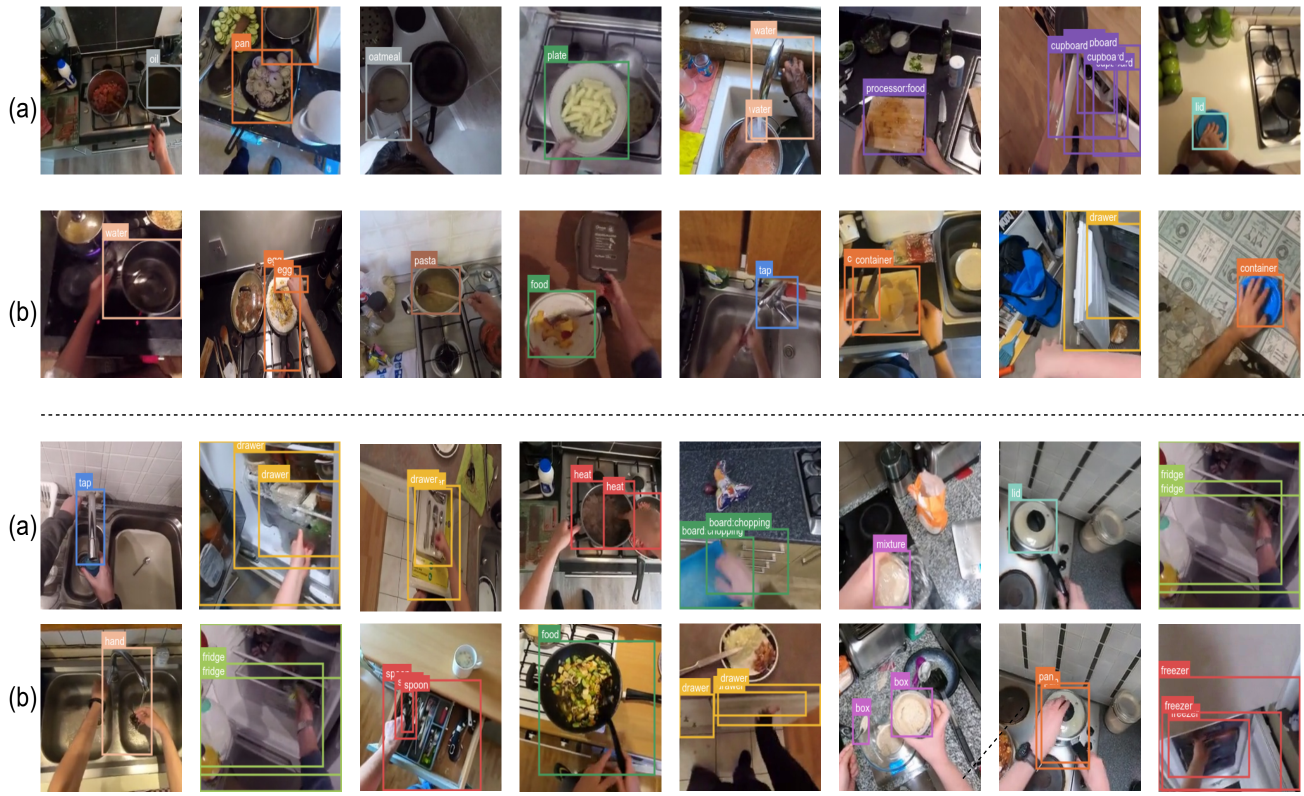

Figure 3. Due to the large number of noun labels in EK-100, similar-looking objects are labeled differently multiple times in the dataset. This confuses our model, NAOGAT since the future action prediction is affected based on the NAO prediction.

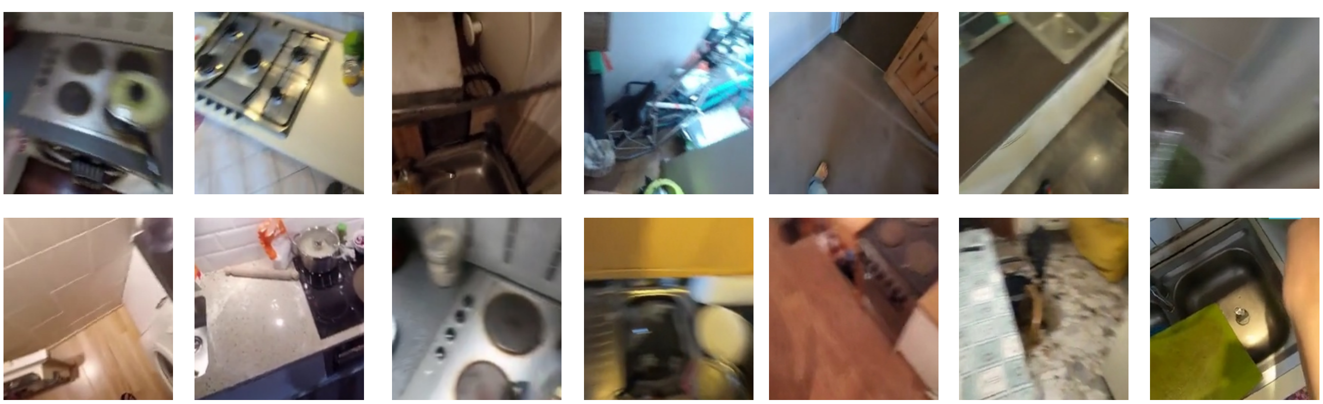

Figure 4. Instances in EpicKitchen-100 where the next-active-object is not detected / not present in the last observed frame.

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[22] Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. Leveraging hand-object interactions in assistive egocentric vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[23] Miao Liu, Siyu Tang, Yin Li, and James Rehg. Forecasting human object interaction: Joint prediction of motor attention and actions in first person video. In *ECCV*, 2020.

[24] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[25] 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

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[26] 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. *arXiv preprint arXiv:2103.14030*, 2021.

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[28] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, June 2022.

[29] Blascovich J.J. Loomis J.M. and A.C. Beall. Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, and Computers*, 31:557–564, 1999.

[30] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1046–1056, 2020.

[31] Tushar Nagarajan and Kristen Grauman. Shaping embodied agent behavior with activity-context priors from egocentric video. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29794–29805. Curran Associates, Inc., 2021.

[32] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE CVPR*, pages 2847–2854, 2012.

[33] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[34] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1569–1578, January 2021.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[36] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer, 2020.

[37] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Anticipating next active objects for egocentric videos, 2023.

[38] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8168–8177, October 2021.

[39] 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 *CVPR*, 2022.

[40] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2020.

[41] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6068–6077, January 2023.