EMAG: Ego-motion Aware and Generalizable 2D Hand Forecasting from Egocentric Videos: Supplementary Materials

Masashi Hatano[1]
hatano1210@keio.jp
Ryo Hachiuma[1,2]
ryo-hachiuma@keio.jp
Hideo Saito[1]
hs@keio.jp

[1] Keio University
Yokohama, Japan
[2] NVIDIA
Taipei, Taiwan

A Dataset Statistics

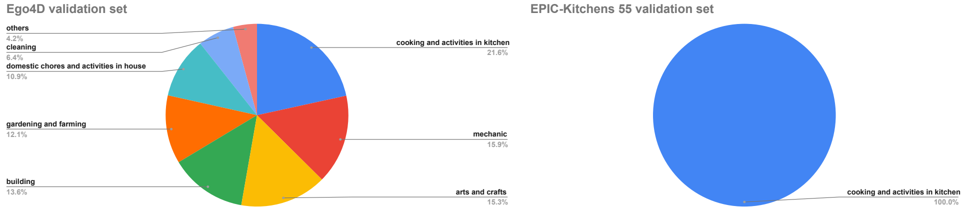

Figure 1: **Scenario breakdown**. The left pie chart represents the scenario breakdown on the validation set of the Ego4D dataset. There are eight categories in total, including inside/outside scenes. The right pie chart represents the scenario breakdown on the validation set of the EPIC-Kitchens 55 dataset. The EPIC-Kitchens 55 dataset contains only one category, cooking and activities in the kitchen.

This section provides statistics on two large-scale egocentric video datasets, Ego4D [4] and EPIC-Kitchens 55 [3]. Fig. 1 presents pie charts illustrating the proportional distribution, categorized by action types or situations, of camera wearers within each validation set of the dataset. The categories are summarized as follows:

- **Cooking and activities in kitchen** contains videos where the camera wearer performs tasks in the kitchen, such as cutting vegetables, washing a pan, and putting dishes away on the shelf.
- **Mechanic** contains situations where the camera wearer uses specific mechanical tools to repair vehicles such as cars or bikes.

- **Arts and crafts** consist of indoor and outdoor scenarios, including activities such as painting and trimming excess materials.
- **Building** category contains a construction scene and a scene depicting brick fabrication.
- **Gardening and farming** consist of both small-scale and large-scale plant caring scenes.
- **Domestic chores and activities in house** contain activities in the house except for the situation in the kitchen, such as laundering, knitting, ironing, and playing cards.
- **Cleaning** category contains cleaning activities such as sweeping with a broom, mopping the floor, and washing a car.
- **Others** consist of various scenarios such as sports (playing basketball or working out at the gym), driving, walking a dog, and activities in the laboratory.

While all videos in the EPIC-Kitchens 55 dataset are categorized as cooking and activities in the kitchen, the Ego4D dataset contains various categories described above. More than three-quarters of the videos in the validation set of Ego4D are composed of cooking and activities in the kitchen (21.6%), mechanic (15.9%), arts/crafts (15.3%), building (13.6%), and gardening/farming (12.1%).

B Implementation Details

Network architecture. We use the dimension size of a token $C = 512$, $k = 2$ for the top-k confidence score with the threshold of 0.5, and set the number of blocks in the encoder and decoder to two. Each block has eight attention heads in the encoder and decoder. Our MLPs for hand and ego-motion prediction consist of a linear layer, an activation function of ReLU [1], a Dropout layer [11], and a final linear layer that outputs the hand positions and ego-motion at future frames.

Optimization. We train the model for 30 epochs using the AdamW optimizer [9], with a peak learning rate of $2e-4$, linearly increased for the first five epochs of the training and decreased to 0.0 until the end of training with cosine decay [8]. We use weight decay of $1e-3$ and a batch size of 64. Regarding the parameters for the loss function, we empirically adapt the control point $\beta = 5.0$, and the balancing loss weight of α is set to one.

C Comparison Methods

We compare with the following methods:

- **CVM** [10]. The Constant Velocity Model (CVM) is a simple but effective trajectory prediction method based on the assumption that the most recent relative motion is the most relevant predictor for the future trajectory. We compute the velocity (v_x, v_y) between $t = T-1$ and $t = T$ for each hand (right, left), and future hand positions for $t = \{T+1, ..., T+F\}$ are forecasted using (v_x, v_y).
- **KF** [6]. The Kalman Filter is an algorithm for estimating a dynamic system's state based on noisy measurements. It tracks the center of the bounding boxes of the hands

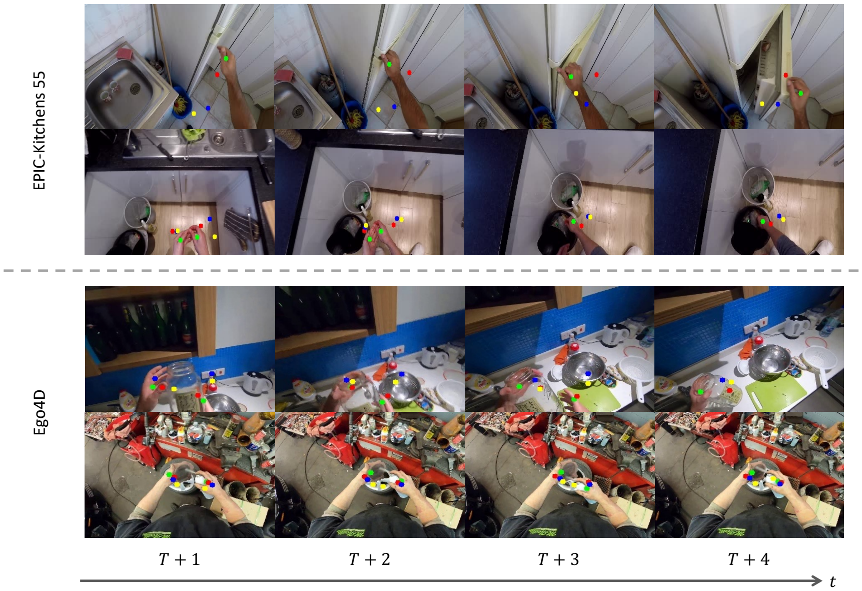

Figure 2: **Qualitative results**. We present two sequences of predictions each from Ego4D and EPIC-Kitchens 55. Dots colored in green, red, blue, and yellow represent the hand positions of the ground truth, the proposed method, I3D + Regression, and OCT, respectively.

with its scale and aspect ratio. Our implementation is based on the code provided by SORT [2][1], which adopts a Kalman Filter to track the center of bounding boxes.

- **Seq2Seq** [12]. Seq2Seq employs Long Short-Term Memory (LSTM) [5] to encode temporal information in the observation sequence and decode the target location of the hands. In our implementation, we adopt the embedding size of 512, the hidden dimension of 256, and the teacher forcing ratio of 0.5 during training.

- **OCT** [7]. OCT simultaneously predicts contact points and the hand trajectory. It takes RGB features extracted by BNInception [13], bounding boxes of hands and objects, and their cropped visual features as input. We modified the model not to predict the contact point for a fair comparison. Our implementation of this model is based on the official implementation[2].

- **I3D + Regression** [4]. This method is proposed as a benchmark for hand forecasting in the Ego4D dataset. The model is trained with the official hand forecasting code[3].

The first two traditional approaches predict based only on past trajectories without training. On the other hand, the last three methods above are recent advanced learning-based approaches in the hand forecasting task.

[1] https://github.com/abewley/sort
[2] https://github.com/stevenlsw/hoi-forecast
[3] https://github.com/EGO4D/forecasting

Table 1: Ablation study of ego-motion representation.

	Ego4D→EPIC	
	ADE ↓	FDE ↓
Background flow	52.08	58.03
Ours	**51.03**	**56.78**

D Further Results

D.1 Qualitative Results

The qualitative results on the Ego4D and EPIC-Kitchens 55 datasets are visualized in Fig. 2. We present two sequences from EPIC-Kitchens 55 in the top two rows of the figure and two sequences from Ego4D in the bottom two rows. In the second sequence from the top of EPIC-Kitchens, where the camera wearer turns left, the proposed method predicts the hand positions more accurately than the other methods. This capability of prediction, even in the presence of ego-motion, verifies the effectiveness of our ego-motion-aware model.

D.2 Ego-motion Representation

We conducted an additional ablation study on ego-motion representation, considering the homography matrix and background optical flow (Tab. 1). The proposed homography matrix representation outperformed the background optical flow representation in cross-scenarios, underscoring the effectiveness of the proposed ego-motion representation.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2016.

[3] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[4] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu,

Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[7] 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.

[8] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[10] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):1696–1703, 2020.

[11] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 15(56):1929–1958, 2014.

[12] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.