Supplementary Materials for: "Generative Adversarial Network for Future Hand Segmentation from Egocentric Video"

Wenqi Jia*, Miao Liu*, and James M. Rehg

Georgia Institute of Technology, Atlanta, United States

This is the supplementary material for our ECCV 2022 paper, titled "Generative Adversarial Network for Future Hand Segmentation from Egocentric Video". The contents are organized as follows.

1 Network Architecture

We present the architecture details of the 3DFCN backbone network (ϕ_E and ϕ_D), Generator Network (G), and Discriminator Network (D) in Table 4. Note that we downsample the head motion flow both spatially and temporally, resulting into a tensor with size $4 \times 7 \times 7 \times 2$. As shown in ID 10 from Table 4, we directly concatenate the generated head motion flow map with the encoded video features and feed the concatenated tensors into the decoder network. Other work [6] proposed to combine optical flow features, learned by a convolutional operation, with video features for future segmentation. However, our empirical finding is that a convolutional flow feature extractor is not necessary, as the decoder can effectively understand the motion pattern incorporated in the low-resolution head motion flow map.

2 Results Using Semantic Masks as Inputs

As discussed in Sec.4.3 from our main paper, the original ConvLSTM [10] and FlowTrans [6] take accurate semantic segmentation masks, from a fine-tuned semantic segmentation model, as inputs. However, the semantic segmentation annotation is not available on the existing egocentric video datasets [7,3] to fine-tune a segmentation model. And the pre-trained Mask-RCNN model obtains

* Equal contribution.

Table 1: *Results using semantic mask as inputs*. Our method outperforms the second-best results (across all methods) by 1.3% on EPIC-Kitchens in average F1 score. The best results are highlighted with **boldface**, and the second-best results are underlined.

Method	EPIC-Kitchens (Precision/ Recall/ F1 Score)		
	short-term	middle-term	long-term
S2S	24.85/ **56.12**/ 34.45	27.59/ **54.69**/ 36.68	26.86/ **52.59**/ 35.55
ConvLSTM	28.24/ 45.48/ 34.84	30.20/ 49.00/ 37.37	29.45/ 47.70/ 36.42
FlowTrans	27.97/ 47.82/ 35.30	29.58/ 52.07/ 37.73	28.95/ 49.89/ 36.64
Ours	**29.09**/ 47.86/ **36.19**	**33.14**/ 47.50/ **39.04**	**32.68**/ 45.05/ **37.88**

sub-optimal results on egocentric video datasets [4]. In this section, we present the results of ConvLSTM, FlowTrans, S2S [1], as well as our method using the semantic segmentation from [4] as inputs. The experimental results are summarized in Table 1. Notably, using the semantic segmentation results from a pre-trained model decreases the model performance on short-term and middle-term hand mask anticipation, yet slightly improves the long-term future hand segmentation results for all methods. More importantly, when using the semantic masks as inputs, our model outperforms the second-best results (across all methods) by 0.9%/1.3%/1.2% in F1 Score for short/middle/long-term future hand segmentation. These results further demonstrate the robustness of our method. It is worth noting that another relevant work from [2] also addresses the future segmentation problem, and can adopt either raw video frames or semantic masks as inputs. However, the official implementation of [2] is under construction. And we found the training of our implementation of [2] to be unstable under the egocentric setting, probably because that the drastic change of scene context incurs additional barriers for the distillation model to generalize.

3 Results Using I3D-Res101 Backbone

We further show our method can generalize to different backbone encoder networks. In Table 2, we report the future hand segmentation results of both our method and 3DFCN baseline using I3DRes50 and I3DRes101 backbone. As discussed in the main paper, the performance improvement of our method (EgoGAN-I3DRes50 vs 3DFCN-I3DRes50) is larger than adopting a denser backbone model (3DFCN-I3DRes101 vs 3DFCN-Res50). Moreover, the EgoGAN model with I3D-Res101 improves 3DFCN-I3DRes101 by +0.1%/0.1%/0.3% on EGTEA and +0.5%/0.4%/0.7% on EPIC-Kitchens. These results further show the robustness of our method. (Note that the performance improvement on EGTEA is relatively small with I3DRes101 backbone, due to the limited training data and dense backbone encoder.)

[1] X2X model described in our main paper was denoted as S2S in [9], when using the semantic mask as inputs

Table 2: *Experimental results using different backbone networks.* Our model achieves consistent performance improvement when using different backbone networks. (See more discussion in Sec. 4)

(a) Experimental Results on EPIC-Kitchens Dataset

Method	Backbone	Epic-Kitchens (Precision/ Recall/ F1 Score)		
		short-term	middle-term	long-term
3DFCN	I3DRes50	69.51/ 70.81/ 70.15	42.51/ 51.66/ 46.64	29.88/ 47.46/ 36.67
	I3DRes101	69.48/ 70.96/ 70.21	42.32/ 52.80/ 46.98	29.97/ 48.37/ 37.01
EgoGAN	I3DRes50	70.89/ 71.24/ 71.07	43.79/ 53.23/ 48.05	31.39/ 48.57/ 38.14
	I3DRes101	69.17/ 74.05/ 71.53	44.09/ 53.79/ 48.46	30.79/ 52.60/ 38.85

(b) Experimental Results on EGTEA Gaze+ Dataset

Method	Backbone	EGTEA (Precision/ Recall/ F1 Score)		
		short-term	middle-term	long-term
3DFCN	I3DRes50	43.62/ 61.69/ 51.11	40.25/ 58.93/ 47.83	37.83/ 58.32/ 45.89
	I3DRes101	44.66/ 61.81/ 51.85	40.49/ 59.72/ 48.26	35.70/ 66.18/ 46.38
EgoGAN	I3DRes50	44.91/ 61.48/ 51.91	41.10/ 59.90/ 48.75	38.16/ 59.88/ 46.61
	I3DRes101	45.69/ 60.42/ 52.03	39.40/ 64.27/ 48.85	36.92/ 64.43/ 46.94

Table 3: *Experimental results on generated future head motion.* We calculate the endpoint error (EPE) between the generated head motion and the ground truth head motion. Our method outperforms HeadReg on the EPIC-Kitchens dataset and works on-par with HeadReg on the EGTEA dataset.

Method	Epic-Kitchens (EPE ↓)	EGTEA (EPE ↓)
HeadReg	10.39	5.27
EgoGAN(Ours)	7.08	5.16

4 Results of Generated Future Head Motion

Our model also has the capability of generating future head motions. In Table 3, we compare our methods with HeadReg – the only baseline model that predicts future head motion. We use the standard endpoint error (EPE) as evaluation metric. On the EPIC-Kitchens dataset, our method outperforms HeadReg by a significant margin. The performance improvement of our method is smaller on the EGTEA dataset, due to fewer available training samples. These results suggest that the GAN from our model can generates more realistic future head motion.

5 Zero-Motion Baseline

We conduct additional experiments to show our method does not simply generate a trivial solution that predicts an "average" hand mask shared across all time steps. Specifically, we consider a zero-motion baseline model that has the same

model design as our method, yet ignores the hand motion. Therefore, the future hand segmentation results $h^{t+\Delta_2}$ and $h^{t+\Delta_3}$ are identical to $h^{t+\Delta_1}$. This baseline model achieves a F1 score of 46.14% and 34.30% for middle-term and long-term future hand segmentation on EPIC-Kitchens, which lags behind our full model (1.9%/3.8% ↓). These results further demonstrate that our method is capable of capturing meaningful hand movements.

6 Pseudo Ground Truth of Hand Masks

Though the domain adaption method from [1] can generate high quality hand segmentation results, they method still yields to the challenging factor of egocentric video, and thus may produce sub-optimal hand segmentation results. Therefore our quantitative experiments cannot reflect the true performance improvement of our method. In Fig. 1, we visualize, hand masks ground truth and prediction results from both our method and FlowTrans. Even though our method demonstrate stronger generalizing capability and predicts more detailed hand shapes, our model has lower F1 score than FlowTrans due to the inaccurate hand masks ground truth.

7 Additional Visualizations

We provide additional visualizations of our results in Fig. 2. Our method can effectively predict future hand masks. However, the model performance drops as the anticipation time increases. This is the same pitfall shared by previous works [8] on visual anticipation. Note that we also provide the video demos of our method.

8 Code and Licenses

The usage of the EPIC-Kitchens Dataset is under the Attribution-NonCommercial 4.0 International License[2]. EGTEA Gaze+ dataset did not provide a license but can be used for research purposes. Our implementation is built on top of [5], which is under the Apache License[3]. Our code will be available at https://github.com/VJWQ/EgoGAN.git.

[2] https://creativecommons.org/licenses/by-nc/4.0/
[3] https://github.com/facebookresearch/SlowFast/blob/main/LICENSE

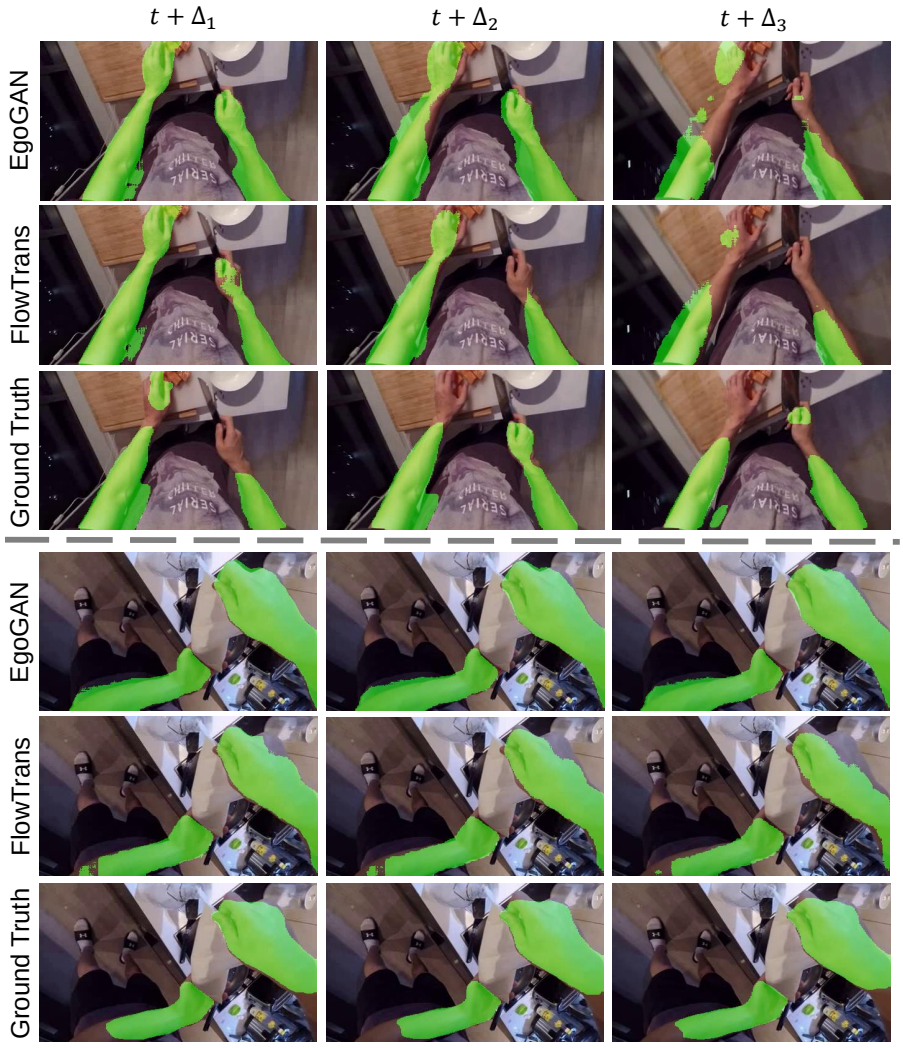

Fig. 1: *Visualization of hand mask ground truth, and prediction results from our model and FlowTrans.* Because of the sub-optimal ground truth, the quantitative results can not demonstrate the true performance improvement of our approach. (See more discussion in Sec. 6)

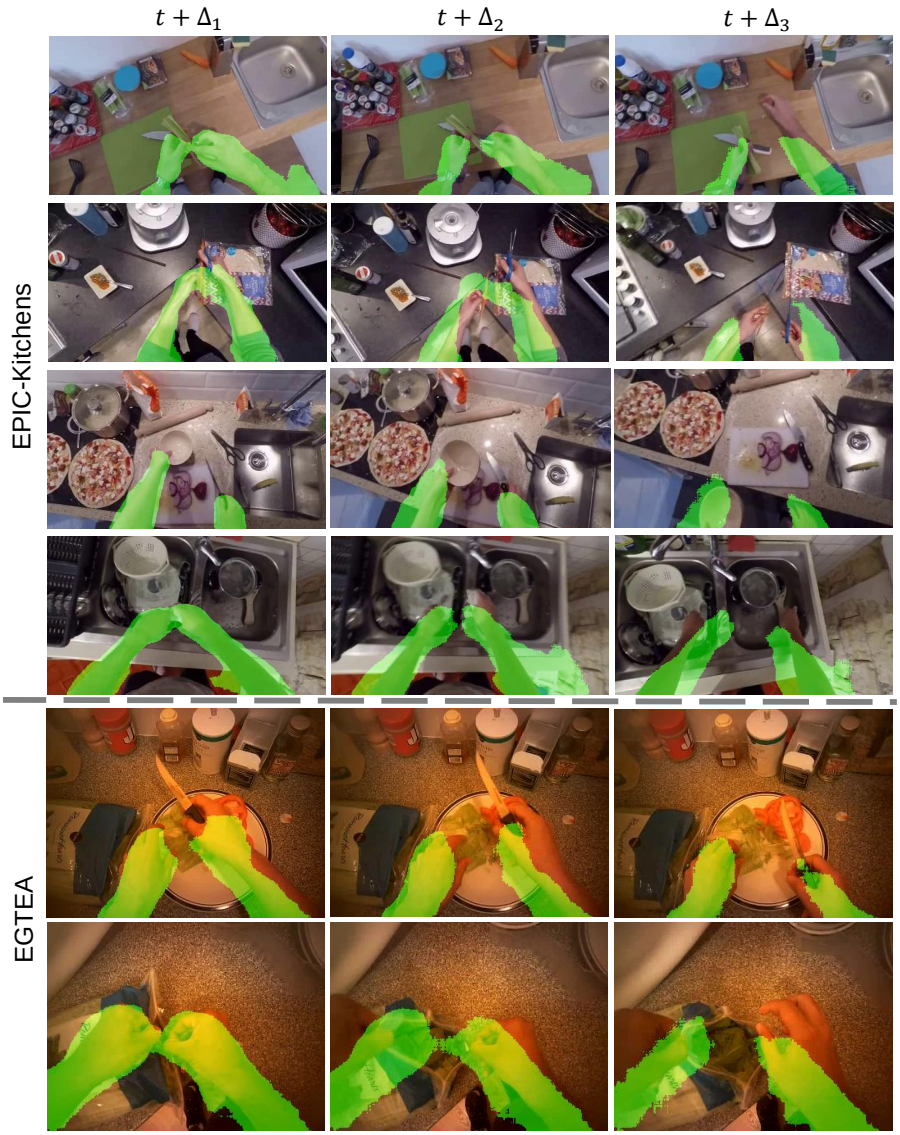

Fig. 2: *Additional Visualization of predicted future hand masks.* From left to right, each column presents the RGB frame at time step t and the short-term $(t+\Delta_1)$, middle-term $(t+\Delta_2)$, and long-term$(t+\Delta_3)$ future hand segmentation results. The first four rows are our results on the EPIC-Kitchens dataset, and the rest three rows are our results on the EGTEA dataset.

Table 4: *Network architecture of our EgoGAN*. We omit the residual connection in backbone I3D-Res50 for simplification. And we present the tensor dimension during training time.

ID	**Branch**	**Type**	**Kernel Size** THW,(C)	**Stride** THW	**Output Size** THWC	**Comments**
1	Encoder Input Size: $8 \times 224 \times 224 \times 3$	Conv3D	$5 \times 7 \times 7, 64$	$1 \times 2 \times 2$	$8 \times 112 \times 112 \times 64$	
2		MaxPool1	$1 \times 3 \times 3$	$1 \times 2 \times 2$	$8 \times 56 \times 56 \times 64$	Skip connect with 24
3		Layer1 Bottleneck 0-2	$3 \times 1 \times 1, 64$ $1 \times 3 \times 3, 64$ (×3) $1 \times 1 \times 1, 256$	$1 \times 1 \times 1$ $1 \times 1 \times 1$ (×3) $1 \times 1 \times 1$	$8 \times 56 \times 56 \times 256$	
4		MaxPool2	$2 \times 1 \times 1$	$2 \times 1 \times 1$	$4 \times 56 \times 56 \times 256$	Skip connect with 23
5		Layer2 Bottleneck 0	$3 \times 1 \times 1, 128$ $1 \times 3 \times 3, 128$ $1 \times 1 \times 1, 512$	$1 \times 1 \times 1$ $1 \times 2 \times 2$ $1 \times 1 \times 1$		
6		Layer2 Bottleneck 1-3	$3 \times 1 \times 1, 128$ $1 \times 3 \times 3, 128$ (×3) $1 \times 1 \times 1, 512$	$1 \times 1 \times 1$ $1 \times 2 \times 2$ (×3) $1 \times 1 \times 1$	$4 \times 28 \times 28 \times 512$	Skip connect with 22
7		Layer3 Bottleneck 0	$3 \times 1 \times 1, 256$ $1 \times 3 \times 3, 256$ $1 \times 1 \times 1, 1024$	$1 \times 1 \times 1$ $1 \times 2 \times 2$ $1 \times 1 \times 1$		
8		Layer3 Bottleneck 1-5	$3 \times 1 \times 1, 256$ $1 \times 3 \times 3, 256$ (×5) $1 \times 1 \times 1, 1024$	$1 \times 1 \times 1$ $1 \times 1 \times 1$ (×5) $1 \times 1 \times 1$	$4 \times 14 \times 14 \times 1024$	Skip connect with 21
9		Layer4 Bottleneck 0	$3 \times 1 \times 1, 128$ $1 \times 3 \times 3, 128$ $1 \times 1 \times 1, 512$	$1 \times 1 \times 1$ $1 \times 2 \times 2$ $1 \times 1 \times 1$		
10		Layer4 Bottleneck 1-2	$3 \times 1 \times 1, 128$ $1 \times 3 \times 3, 128$ (×2) $1 \times 1 \times 1, 512$	$1 \times 1 \times 1$ $1 \times 2 \times 2$ (×2) $1 \times 1 \times 1$	$4 \times 7 \times 7 \times 2048$	Concat with generated future head motion
11	Generator Network (G) Input Size: $4 \times 7 \times 7 \times 2048$	Conv3d 1	$1 \times 1 \times 1, 2048$	$1 \times 1 \times 1$	$4 \times 7 \times 7 \times 1024$	
12		Conv3d 2	$1 \times 1 \times 1, 1024$	$1 \times 1 \times 1$	$4 \times 7 \times 7 \times 512$	
13		Conv3d 3	$1 \times 1 \times 1, 512$	$1 \times 1 \times 1$	$4 \times 7 \times 7 \times 2$	
14		Tanh			$4 \times 7 \times 7 \times 2$	Input for Decoder & Discriminator
15	Discriminator Network (D) Input Size: $4 \times 7 \times 7 \times 2$	Conv3d 1	$1 \times 3 \times 3, 2$	$1 \times 1 \times 1$	$4 \times 5 \times 5 \times 32$	
16		Conv3d 2	$1 \times 3 \times 3, 32$	$1 \times 1 \times 1$	$4 \times 3 \times 3 \times 64$	
17		Conv3d 3	$1 \times 3 \times 3, 64$	$1 \times 1 \times 1$	$2 \times 1 \times 1 \times 128$	
18		Adaptive Avg Pooling			$1 \times 1 \times 1 \times 128$	
19		Linear			1	
20		Sigmoid			1	BCE loss
21	Decoder Input Size: $4 \times 7 \times 7 \times 2050$	ConvTranspose3d 1	$1 \times 3 \times 3, 2050$	$1 \times 2 \times 2$	$4 \times 14 \times 14 \times 1024$	Skip connect with 8
22		ConvTranspose3d 2	$1 \times 3 \times 3, 1024$	$1 \times 2 \times 2$	$4 \times 28 \times 28 \times 512$	Skip connect with 6
23		ConvTranspose3d 3	$1 \times 3 \times 3, 512$	$1 \times 2 \times 2$	$4 \times 56 \times 56 \times 256$	Skip connect with 4
24		ConvTranspose3d 4	$3 \times 3 \times 3, 256$	$1 \times 1 \times 1$	$8 \times 56 \times 56 \times 64$	Skip connect with 2
25		ConvTranspose3d 5	$1 \times 5 \times 5, 64$	$1 \times 4 \times 4$	$8 \times 224 \times 224 \times 64$	
26		Conv3d 1	$4 \times 1 \times 1, 64$	$1 \times 1 \times 1$	$5 \times 224 \times 224 \times 32$	
27		Conv3d 2	$3 \times 1 \times 1, 32$	$1 \times 1 \times 1$	$3 \times 224 \times 224 \times 16$	
28		Conv3d 3 (Classifier)	$1 \times 1 \times 1, 16$	$1 \times 1 \times 1$	$3 \times 224 \times 224 \times 1$	BCE loss

References

1. Cai, M., Lu, F., Sato, Y.: Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In: CVPR (2020) 4
2. Chiu, H.k., Adeli, E., Niebles, J.C.: Segmenting the future. ICRA-L (2020) 2
3. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018) 1
4. 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) 2
5. Fan, H., Li, Y., Xiong, B., Lo, W.Y., Feichtenhofer, C.: Pyslowfast. https://github.com/facebookresearch/slowfast (2020) 4
6. Jin, X., Xiao, H., Shen, X., Yang, J., Lin, Z., Chen, Y., Jie, Z., Feng, J., Yan, S.: Predicting scene parsing and motion dynamics in the future. In: NeurIPS (2017) 1
7. Li, Y., Liu, M., Rehg, J.M.: In the eye of the beholder: Gaze and actions in first person video. TPAMI (2021) 1
8. Liu, M., Tang, S., Li, Y., Rehg, J.: Forecasting human object interaction: Joint prediction of motor attention and actions in first person video. In: ECCV (2020) 4
9. Luc, P., Neverova, N., Couprie, C., Verbeek, J., LeCun, Y.: Predicting deeper into the future of semantic segmentation. In: ICCV (2017) 2
10. Rochan, M., et al.: Future semantic segmentation with convolutional lstm. In: BMVC (2018) 1