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

Abstract. We introduce the novel problem of anticipating a time series of future hand masks from egocentric video. A key challenge is to model the stochasticity of future head motions, which globally impact the head-worn camera video analysis. To this end, we propose a novel deep generative model – EgoGAN. Our model first utilizes a 3D Fully Convolutional Network to learn a spatio-temporal video representation for pixel-wise visual anticipation. It then generates future head motion using the Generative Adversarial Network (GAN), and predicts the future hand masks based on both the encoded video representation and the generated future head motion. We evaluate our method on both the EPIC-Kitchens and the EGTEA Gaze+ datasets. We conduct detailed ablation studies to validate the design choices of our approach. Furthermore, we compare our method with previous state-of-the-art methods on future image segmentation and provide extensive analysis to show that our method can more accurately predict future hand masks. Project page: https://vjwq.github.io/EgoGAN/

Keywords: Egocentric Vision, Hand Segmentation, Visual Anticipation

1 Introduction

The egocentric vision paradigm provides an ideal vehicle for studying the relationship between visual anticipation and intentional motor behaviors, as head-worn cameras can capture both human visual experience and related sensory-motor signals. While prior works have recently addressed action anticipation in an egocentric setting [11,24,32,49,13], the problem of forecasting the detailed shape of hand movements in egocentric video remains unexplored. This is a significant deficit because many everyday motor behaviors cannot be easily categorized into specific action classes and yet play an important role in preparing and executing our routine activities. Such a general prediction capability could enable new applications in Augmented Reality (AR) and robotics, such as monitoring for safety in dangerous environments such as construction sites, or facilitating human-robot collaboration via improved anticipation.

To bridge this gap, this paper introduces a novel task of forecasting the detailed representation of future hand movements in egocentric video. Specifically,

* Equal contribution.

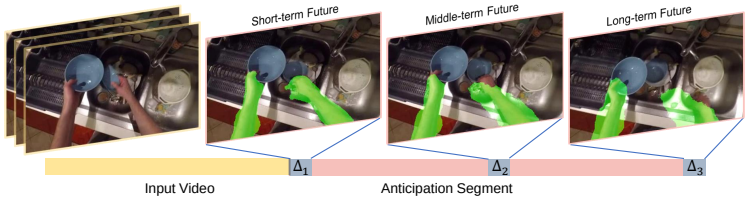

Fig. 1: *Future hand segmentation task*: Given an input egocentric video, our goal is to predict a time series of future hand masks in the anticipation video segment. Δ_1, Δ_2, and Δ_3 represent the short-term, middle-term, and long-term time points in the anticipation segment, respectively. The entanglement between drastic head motion and non-rigid hand movements poses a significant technical barrier in computer vision. Here, we visualize our forecasting results on this challenging task (best viewed in color).

given an egocentric video, we seek to predict the hand masks of future video frames at three time points defined as short-term, middle-term, and long-term future (see Fig. 1 for a visual illustration of our problem setting). This task is extremely challenging for two reasons: 1) hands are deformable and capable of fast movement, and 2) head and hand motion are entangled in the egocentric video. Addressing these challenges requires the ability to 1) address the inherent uncertainty in anticipating the no-rigid hand movements, and 2) explicitly model the coordination between head and hand [42].

We attack the unique challenges of hand segmentation prediction by introducing a novel deep model – *EgoGAN*. Our model adopts a 3D Fully Convolutional Network (3DFCN) as the backbone to learn spatio-temporal video features. We then utilize the Generative Adversarial Network (GAN) to aid pixel-wise visual anticipation. Instead of using GAN to directly generate future video frame pixels from egocentric videos as in [67], our key insight is to use the GAN to model an underlying distribution of possible future head motion. The adopted generative adversarial training schema can account for the uncertainty of future hand movements anticipation. In addition, the generated future head motion provides ancillary cues that complement video features for anticipating complex egocentric hand movements. Our end-to-end trainable EgoGAN model uses future hand masks as supervisory signals to train the segmentation network and estimated sparse optical flow maps from head motions to train the Generator and the Discriminator. At inference time, our model predicts a time series of future hand masks based *only* on the egocentric video frame inputs.

To demonstrate the benefits of our proposed EgoGAN, we evaluate our model on two egocentric video datasets: EPIC-Kitchens 55 [5] and EGTEA Gaze+ [28]. We first conduct detailed ablation studies to validate our model design, and then compare our approach to the state-of-the-art methods on future image segmentation, demonstrating consistent performance gains on both datasets. We further provide visualizations to show the effect of our method. In summary, our paper makes following contributions:

- We introduce a novel problem of predicting a time series of future hand masks from egocentric videos.
- We propose a novel deep generative model – EgoGAN, that hallucinates future head motions and further predicts future hand masks. To the best of our knowledge, we are the first to use a GAN to generate egocentric motion cues for visual anticipation.
- We conduct comprehensive experiments on two benchmark egocentric video datasets: EPIC-Kitchens 55 [5] and EGTEA Gaze+ [29]. Our model achieves 1.3% performance improvements on EPIC-Kitchens and 0.7% on EGTEA in average F1 score. We also provide visualizations of our results and additional discussion of our method.

2 Related Work

We first review the most relevant works on egocentric vision. We then discuss previous literature on future image segmentation. Furthermore, we describe the related efforts on developing generative models for visual anticipation.

Hands in Egocentric Vision. Previous efforts on egocentric vision addressed a variety interesting problems, including action analysis [45,8,28,31,44,11,32,53,24,38] and social interaction understanding [7,52,65,63], etc. Here, we focus on discussing prior works on learning hand representations from egocentric videos. The most relevant work is from Liu et al. [32], where they factorized the future hand positions a latent attentional representation for action anticipation without considering the head motion. Similarly, Dessalene et al. [6] focused on predicting the hand-object interaction region of an action. Fathi et al. [9] utilized hand-eye coordination to design a probabilistic model for gaze estimation. Li et al. [30] showed how the motion patterns of the hands can be utilized for egocentric action recognition. Ma et al. [36] made use of a hand segmentation network to predict hand masks for localizing the object of interest and further recognizing the action. Shen et al. [49] proposed to use hand mask and gaze fixation as additional cues for action anticipation. Rather than anticipating the hand movements, these previous works mainly use egocentric hand movements as an additional modality or intermediate representation for egocentric action understanding. Recently, Cai et al. [1] proposed a Bayesian-based domain adaptation framework for hand segmentation on egocentric video frames. In contrast, we address the novel task of predicting pixel-wise hand masks, which captures the fine-grained details of future hand movements.

Future Segmentation. A rich set of literature addressed the related but vastly different task of video segmentation [64,54,62,3,39]. We refer to a recent survey [60] for a thorough discussion on this topic. Note that previous works on video segmentation seek to track the instance masks within the video segment, and therefore do not apply to the anticipation setting, where the information of future video frames is not accessible for making an inference. Fewer works address the more relevant topic of future image semantic segmentation. Luc et al. [35] first investigated the problem of semantic segmentation of future video

frames and further extended their work to future instance segmentation [34]. Nabavi et al. [46] utilized the ConvLSTM network to model the temporal correlations of video sequences for future semantic segmentation. Jin et al. [22] proposed to anticipate the future optical flow and future scene segmentation jointly. Recently, Chiu et al. [4] introduced a teacher-student knowledge distillation model for predicting the future semantic segmentation based on preceding video frames. Building on these prior works, we propose the first model to address the future segmentation problem under the challenging egocentric setting. It is worth noting that previous methods recursively predict future segmentation, in which the current anticipation result is used as the input for predicting the segmentation of the next time step. In contrast, we use a 3D Fully Convolutional Network (3DFCN) to predict a time series of future hand masks in one shot. In Sec. 4.3, we show that the 3DFCN can effectively capture the spatio-temporal video features for pixel-wise visual anticipation in an end-to-end fashion. We also compare our EgoGAN model to those relevant works and demonstrate a clear performance improvement.

Generative Models for Visual Anticipation. Tremendous efforts have been made in action anticipation [25,56,12,23,11,32,53,24,47,16] and generative adversarial networks [15,40,66,14,37,21]. Here we mainly discuss previous investigations on forecasting the human body motions using generative models. Fragkiadaki et al. [10] proposed to use a recurrent network for predicting and generating the human body poses and dynamics from videos. A similar idea was also explored in [17]. Walker et al. [58] utilized Variational Autoencoders (VAE) for predicting the dense trajectories of video pixels. They further leveraged human body poses as an intermediate feature for generating future video frames with a Generative Adversarial Network (GAN) [59]. Gupta et al. [18] explored a GAN-based model for forecasting human trajectories. Zhang et al. [69,68] developed a Conditional Variational Autoencoder to generate human body meshes and motions in 3D scenes. Despite the success in forecasting body motion, the use of GANs was largely understudied in egocentric vision. Zhang et al. [67] used a GAN to generate future video frames and further predict future gaze fixation. Though GAN has the capability of addressing the uncertainty of data distribution, using GANs to directly forecast pixels in video [55] remains a challenge, especially when there exists drastic background motion in the egocentric videos [67]. In contrast, our method adopts the adversarial training mechanism to model the underlying distribution of possible future head motion, and thereby captures the drastic change of scene context in egocentric video. In the ablation study, we show that our approach outperforms a baseline model that uses GAN to directly predict future hand masks.

3 Method

Given an input egocentric video $x = \{x^1, ..., x^t\}$, where x^t is the video frame indexed by time t, our goal is to predict a time series of future hand masks $h = \{h^{t+\Delta_1}, h^{t+\Delta_2}, h^{t+\Delta_3}\}$. As illustrated in Fig. 1, we consider hand segmentation

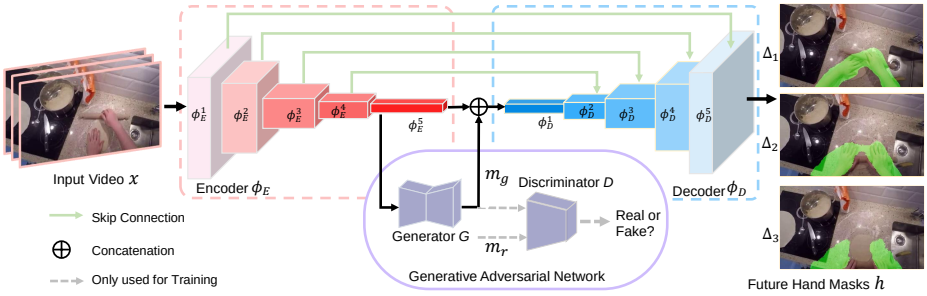

Fig. 2: *Overview of our proposed EgoGAN model.* Our model takes egocentric video frames as the inputs, and outputs future hand masks at different time steps. It is composed of a **3D Fully Convolutional Network (3DFCN)** and a **Generative Adversarial Network (GAN)**. The Encoder Network ϕ_E in the **3DFCN** extracts video features from the input frames, and is then separated into two branches: (1) encoded feature $\phi_E(x)$ is fed into the **Generator (G)** in **GAN** for generating fake future head motion m_g, and a **Discriminator (D)** is trained to distinguish the generated future head motion from the real ones; (2) m_g is concatenated to $\phi_E(x)$ and the concatenated tensor are then fed into the Decoder Network ϕ_D in **3DFCN**. Finally, the encoder features are further combined with corresponding decoder features using skip connections for future hand mask prediction.

as a binary classification problem: the value of $h^i(x, y)$ can be viewed as the probability of spatial position (x, y) being a hand pixel at time step i, where $i \in \{t + \Delta_1, t + \Delta_2, t + \Delta_3\}$. Δ_1, Δ_2, and Δ_3 represent the time steps for short-term, middle-term, and long-term future segmentation, respectively. This three-steps-ahead visual anticipation setting is also used in previous works on future image segmentation [35,46].

We now present an overview of our **EgoGAN** model in Fig. 2. We make use of a 3D Fully Convolutional Network (3DFCN) ϕ as the backbone model for future hand segmentation. The 3DFCN is composed of a 3D convolutional encoder ϕ_E and a 3D deconvolutional decoder ϕ_D. We further adopt a Generative Adversarial Network (GAN) for learning future head motions. Specifically, a Generator network (G), composed of 3D convolutional operations, is used to generate future head motion m_g based on the encoded video feature $\phi_E(x)$. A Discriminator Network (D) is trained to distinguish the fake future head motions m_g from real future head motions m_r. Finally, ϕ_D combines m_g and $\phi_E(x)$ for predicting future hand masks. In the following sections, we detail each key component of our model.

3.1 3D Fully Convolutional Network

We first introduce the 3D Fully Convolutional Network (3DFCN) backbone in our method. We use an I3D model [61] as the backbone encoder network ϕ_E for learning spatio-temporal video representations. Following [51,19], ϕ_E has 5

convolutional blocks, thereby producing video features at different spatial and temporal resolutions. Following [33], we construct the decoder network ϕ_D symmetric to ϕ_E. Therefore, ϕ_D is also composed of 5 deconvolution layers. We denote the encoder and decoder video features from the ith convolutional block as $\phi_E^i(x)$ and $\phi_D^i(x)$, respectively (See Fig. 2 for the index naming of ϕ_E and ϕ_D). The features of each decoder layer are combined with the features from the corresponding encoder block with skip connections and are then fed into the next layer. Formally, we have:

$$\phi_D^{i+1}(x) = deconv(\phi_D^i(x) + \phi_E^{6-i}(x)), \tag{1}$$

where $i \in \{1, 2, 3, 4\}$. We design our decoder so that ϕ_D^i produces a feature map with the same tensor size as $\phi_E^{6-i}(x)$. The deconvolution operation is implemented with 3D transposed convolution. Note that the last deconvolution layer of ϕ_D produces a tensor of the same size as the input video ($T \times W \times H$). We further apply a 3D convolutional operation with a kernel size of $k \times 1 \times 1$ to predict the future hand mask tensor h with size $3 \times W \times H$, where each temporal slice corresponds to the predicted hand masks of the short-term, middle-term, and long-term future video frames. We describe the details of our network architecture in the supplementary materials.

3.2 Generative Adversarial Network

The key to our approach is to use the Generative Adversarial Network (GAN) to hallucinate the future head motions for future hand mask segmentation. Our design choice stems from the observation that head motion causes drastic changes in the active object cues and background scene context captured in the egocentric videos, and this motion is closely related to hand movements. Therefore, we seek to explicitly encode the future head motion cues for hand motion anticipation. Moreover, visual anticipation has intrinsic ambiguity – similar current observations may correspond to different future outcomes. This observation motivates us to use the adversarial training scheme to account for the inherent uncertainty of future representation. In this section, we introduce the egocentric head motion representation. We then describe the design choice and learning objective of the GAN in our method.

Egocentric Head Motion Representation. In the egocentric setting, head motion is implicitly incorporated in the video itself. Thus, we follow [27] to use the sparsely sampled optical flow to represent the egocentric head motion. As mentioned before, the real future head motion is denoted as m_r, and is only available for training.

Generator Network and Discriminator Network. The generator network (G) takes video feature $\phi_E(x)$ as inputs and generates future head motions $m_g = G(\phi_E(x))$. Following [57,67,21,59], G does not take any noise variables as additional inputs. This is because the $\phi_E(x)$ is a latent representation that incorporates the noisy signals of visual anticipation. G is composed of multiple 3D convolutional operations and a nonlinearity function, and is trained to

produce a realistic m_g that is difficult to distinguish from m_r for an adversarially-trained discriminator network (D). D takes future head motion samples as inputs and determines whether the input sample is real or fake. It is composed of 3D convolutional operations and a sigmoid function for binary classification, and is trained to classify the input sample as either real or generated.

Learning Objective of GAN. We now formally define the objective function of the GAN in our method. The objective function for training the discriminator network is given by:

$$\mathcal{L}_d = \mathcal{L}_{ce}(D(m_r), 1) + \mathcal{L}_{ce}(D(m_g), 0), \tag{2}$$

where $\mathcal{L}_{ce}$ is the standard cross-entropy loss for binary classification. The generator loss $\mathcal{L}_g$ can be formulated as:

$$\mathcal{L}_g = \mathcal{L}_{ce}(D(m_g), 1) + \lambda |m_g - m_r|. \tag{3}$$

Here, we follow [41] to adopt a traditional L1 distance loss that encourages the generated sample to be visually consistent with the real sample, while λ denotes the weight to balance the two loss terms.

3.3 Full Model of EgoGAN

We now summarize the full architecture of our proposed EgoGAN model. The main idea is to explicitly model the underlying distribution of possible future head motion m_g with the GAN, and use m_g as additional cues to facilitate future hand mask segmentation from the video representations of the encoder network. Specifically, the video feature from the last encoder block $\phi_E^5(x)$ and generated future head motions m_g are concatenated and fed into the first layer of the decoder as inputs. Therefore, we have:

$$\phi_D^1(x) = deconv(\phi_E^5(x) \oplus m_g). \tag{4}$$

Hence, the decoder network jointly considers $\phi_E(x)$ and m_g for predicting future hand masks h.

Training and Inference. We use the binary cross-entropy loss to train the 3DFCN encoder and decoder:

$$\mathcal{L}_{seg} = \mathcal{L}_{ce}(\phi_D(\phi_E(x), m_g), \hat{h}), \tag{5}$$

where $\hat{h}$ denotes the ground truth of future hand masks. We adopt the standard adversarial training pipeline in [14], where G and D are trained to play against each other. Therefore, we let the gradients alternatively flow through D, and then G. Moreover, we freeze the encoder weights during the gradient step on G and D, and freeze the generator weights during the gradient step on the 3DFCN to isolate their training processes from each other.

Note that our model does not need the real future head motion as additional inputs at inference time. Instead, our model can generate future head motion and further predict future hand masks based on only raw video frames.

3.4 Implementation Details

Network Architecture. We adopt an I3D-Res50 model [2,61] that is pre-trained on Kinetics as the backbone encoder network. It is composed of five 3D convolutional blocks, connecting with a symmetrical decoder network that contains five 3D deconvolutional layers. As for the GAN network, the generator network takes the video features from the 5th block of the encoder network as inputs and produces a low-resolution future head motion flow map as output. The discriminator network serves as a binary classifier to supervise the quality of the output of the generator. Our model is implemented in PyTorch and will be made publicly available.

Training Schema. As discussed in the previous section, the gradients step separately for 3D Fully Convolutional Network (3DFCN), Generator (G), and Discriminator (D). The 3DFCN model is trained using an SGD optimizer with momentum of 0.9. The initial learning rate is 0.1 with cosine decay. We set weight decay to 1e-4 and enable batch norm [20]. G and D are trained using the Adam Optimizer with momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a initial learning rate of 0.01 with cosine decay. Our model was trained for 70 epochs with batch size 16 on 4 GPUs, and synchronized batch normalization was enabled.

Data Processing. We downsampled all video frames to a height of 256 while preserving the original aspect ratio. For training, we applied several data augmentation techniques, including random flipping, rotation, cropping, and color jittering to avoid overfitting. Our model takes an input of 8 frames (temporally sampled by 8) with a resolution of 224×224. We use the TV-L1 algorithm [43] to compute the optical flow, and sparsely sample from the computed flow map to approximate the head motion as discussed in Sec. 3.2. Therefore, the head motion is represented as a sparse flow map spatially downsampled by 32. At inference time, our model takes the downsampled videos with the original aspect ratio as inputs, and predicts the future hand masks.

4 Experiments

4.1 Dataset and Metrics

Dataset. We make use of two egocentric video benchmark datasets: EPIC-Kitchens 55 [5] and EGTEA Gaze+ [29]. For the EPIC-Kitchens dataset, we set $\delta_{1,2,3} = \{1, 15, 30\}$, which corresponds to a long-term anticipation time of $1.0s$. As for the EGTEA dataset, we set $\Delta_{1,2,3} = \{1, 6, 12\}$, which corresponds to an anticipation time of $0.5s$, because EGTEA has a smaller angle of view in comparison with the EPIC-Kitchens. The same anticipation time setup is also adopted in [32]. To encourage our model to capture the meaningful preparation and planning process of daily actions, we segment the data so that the long-term future frame is chosen right before the beginning of each trimmed action segment annotated in EPIC-Kitchens and EGTEA. We use the train/val split provided by [11] for EPIC-Kitchens 55 and the train/test split1 from EGTEA. We remove the instances where hands are not captured within the anticipation

segment, which results in 11,935/2,746 (train/val) samples on EPIC-Kitchens, and 4,042/991 (train/test) samples on EGTEA.

Hand Mask Ground Truth. For the EPIC-Kitchens dataset, we use the domain adaption method introduced in [1] to generate the ground truth hand masks. [1] has empirically verified the quality of generated hand masks. As for the EGTEA dataset, we train a 2D FCN model for frame-level hand segmentation using the provided hand mask annotation. As discussed in [26], the FCN model can generalize well on the entire dataset. We thus use the inference results on the anticipation video frames as the ground truth of future hand masks.

Metrics. As discussed in Sec. 3, we consider future hand segmentation as a pixel-wise binary classification problem. Previous future image segmentation works [4,22] use pixel accuracy and mIoU as evaluation metrics. However, pixel accuracy does not penalize the false-negative prediction of the long-tailed distribution, and mIoU can not properly evaluate the shape of the predicted masks for binary segmentation. Therefore, we follow [28,32] to report Precision and Recall values together with their corresponding F1 scores.

4.2 Model Ablations and Analysis

To validate our model design, we conduct experiments on ablations and variations in our model. Specifically, we investigate how the egocentric head motion cues facilitate future hand segmentation and demonstrate the benefits of using the GAN for modeling future head motion. We also show how modeling the future gaze as attentional representation affects the future hand segmentation performance.

Benefits of Encoding Future Head Motions. As a starting point, we compare the model that uses only the 3D Fully Convolutional Network (denoted as *3DFCN*) with the model that directly takes future head motion as an additional input modality (denoted as *HeadDir*). HeadDir shares the same backbone network as 3DFCN, but requires the future head motions for making an inference and therefore violates the future anticipation setting, where the model can not use any information from the anticipation video segment for making an inference. HeadDir quantifies the performance improvement when the egocentric head motion cues are explicitly encoded into the model in a two-stream structure [50]. The experimental results are summarized in Table 1. Compared to 3DFCN, HeadDir achieves a large performance gain on EPIC-Kitchens (+0.8%/1.2/1.1% in F1 score for short/middle/long term anticipation), and reaches (+1.4%/1.3%/1.3%) on EGTEA.

Our method, on the other hand, consistently outperforms 3DFCN on both EPIC-Kitchens(+0.9%/1.5%/1.8%) and EGTEA (+0.8%/0.9%/0.7%). More importantly, our method improves HeadDir by +0.1%/0.2%/0.4% on EPIC-Kitchens. This result suggests that the GAN from our model does not simply learn to predict a future head motion flow map; instead, it models the underlying distribution of possible future head motion and thus improves the future hand anticipation accuracy by addressing the inherent uncertainty of visual forecasting. It is to be observed that our model slightly lags behind HeadDir (0.6%/0.4%/0.6% $\downarrow$)

Table 1: *Analysis of variations in our approach.* We conduct detailed ablation studies to validate our model design, and further show the results of variations of our method to demonstrate the benefits of using the GAN for modeling future head motion. *: HeadDir takes future head motions as additional input modalities at inference time, which in fact violates the future anticipation setting (See more discussion in Sec. 4.2). The best results are highlighted with **boldface**.

(a) Experimental Results on EPIC-Kitchens Dataset

Method	EPIC-Kitchens (Precision/ Recall/ F1 Score)		
	short-term	middle-term	long-term
Future Gaze	N/A	N/A	N/A
HeadDir*	70.55/ 71.33/ 70.94	43.15/ 53.66/ 47.83	30.51/ 49.60/ 37.78
3DFCN (w/o GAN, w/o Head)	69.51/ 70.81/ 70.15	42.51/ 51.66/ 46.64	29.88/ 47.46/ 36.67
HeadReg (w/o GAN, w/ Head)	70.46/ 70.25/ 70.36	41.41/ 52.55/ 46.32	29.22/ 48.50/ 36.47
DirectGan (w/ GAN, w/o Head)	69.12/ **71.60**/ 70.34	43.83/ 51.32/ 47.28	30.76/ 47.48/ 37.33
EgoGAN (w/ GAN, w/ Head)	**70.89**/ 71.24/ **71.07**	**43.79**/ **53.23**/ **48.05**	**31.39**/ **48.57**/ **38.14**

(b) Experimental Results on EGTEA Gaze+ Dataset

Method	EGTEA (Precision/ Recall/ F1 Score)		
	short-term	middle-term	long-term
Future Gaze	45.17/ 59.94/ 51.51	38.63/ 64.02/ 48.19	35.71/ 63.78/ 45.78
HeadDir*	44.58/ 63.87/ 52.51	41.29/ 60.65/ 49.13	39.36/ 59.02/ 47.23
3DFCN (w/o GAN, w/o Head)	43.62/ **61.69**/ 51.11	40.25/ 58.93/ 47.83	37.83/ 58.32/ 45.89
HeadReg (w/o GAN, w/ Head)	43.54/ 61.03/ 50.82	**41.31**/ 55.24/ 47.27	36.87/ 58.23/ 45.15
DirectGan (w/ GAN, w/o Head)	43.78/ 61.33/ 51.09	38.38/ **63.81**/ 47.93	35.53/ **63.41**/ 45.54
EgoGAN (w/ GAN, w/ Head)	**44.91**/ 61.48/ **51.91**	41.10/ 59.90/ **48.75**	**38.16**/ 59.88/ **46.61**

on EGTEA, because EGTEA has fewer samples to train our deep generative model. And we also re-emphasize that our method does not use any additional inputs at inference time as in HeadDir.

The Effect of GAN. To further show the benefits of using the GAN for learning future head motions, we consider a baseline model – *HeadReg*, that uses a regression network to predict future head motions with only L1 distance in Eq. 3. Note that the regression network is implemented the same way as the generator network from EgoGAN. As shown in Table 1, without using an adversarial training mechanism in our approach, HeadReg lags behind our model by 0.7%/1.7%/1.7% ↓ and 1.1%/1.5%/1.5% ↓ in F1 score for short/middle/long term anticipation on EPIC-Kitchens and EGTEA, respectively. These results support our claim that the GAN can address the stochastic nature of representation and thereby outperforms HeadReg by a notable margin on the future hand segmentation task.

Video Pixel Generation vs. Head Motion Generation. We denote another baseline model that directly uses a GAN for anticipating future hand masks, as *DirectGan*. This model is composed of the 3DFCN backbone network that generates the future hand masks, and a discriminator network that classifies whether the given hand masks are real or not. The results are presented in Table 1. Importantly, the adversarial training schema in DirectGan slightly

decreases the performance of 3DFCN model on EGTEA, and has minor improvement on EPIC-Kitchens. We speculate that this is because directly using a GAN for predicting future hand masks cannot effectively capture the drastic change of scene context in egocentric video. In contrast, our model uses a GAN to explicitly model the head-hand coordination in the egocentric video thereby being capable of more accurately forecasting egocentric hand masks.

Future Head Motion vs. Future Gaze. Furthermore, we present experimental results on how modeling future gaze fixation affects future hand segmentation. Note that the gaze tracking data is only available for the EGTEA dataset. Specifically, we make use of a GAN to model the probabilistic distribution of future gaze fixation. Instead of concatenating future gaze with encoded video features as in Eq. 4, we follow [28] to use gaze distribution as a saliency map to select important spatio-temporal video features with element-wise multiplication. As shown in Table 1, the resulting future gaze model slightly outperforms the baseline 3DFCN model, yet lags behind our model that uses head motion as the key representation (0.7%/0.6%/0.6% $\downarrow$ in F1 score on EGTEA). Previous work [27] suggested that eye-head-hand coordination is important for egocentric gaze estimation, while our results further show that exploiting the eye-head-hand coordination is also beneficial for pixel-wise egocentric visual anticipation. Moreover, future head motion potentially plays a more important role than future gaze fixation on our fine-grained hand forecasting task.

Analysis on Ablation Studies. To help interpret the performance improvement of our method, we consider a baseline 3DFCN model that uses dense I3D-Res101 as the encoder network. Importantly, with 50 more layers, the I3D-Res101 backbone can only improve the model performance by +0.1%/0.3%/0.3% on EPIC-Kitchens and +0.7%/0.4%/0.5% on EGTEA. As shown in Table 1, our model has a larger performance improvement than switching to a dense encoder network. In supplementary material, we also present additional results of our model using the I3D-Res101 backbone and further demonstrate our method is a robust approach that can generalize to different backbone networks.

4.3 Comparison to State-of-the-Art Methods

We are the first to address the challenging problem of future hand segmentation from the egocentric video. We note that another branch of prior work considered the related problem of future image segmentation [64,54,62,3,39], track instances masks over time, and therefore can not be used to address the future segmentation problem where the future video frames are not available as inputs for the tracking model. Therefore, we adapt previous state-of-the-art future image segmentation methods to our problem setting and consider the following strong baselines (additional discussion of the baseline choices can be found in the supplementary material):

•X2X [35] proposes a recursive method that uses the anticipated mask at time step $t+1$ as an input to predict the future masks at time step $t+2$, and so forth.

•FlowTrans [22] jointly predicts the masks and optical flow at time step $t+1$ and recursively predicts the future masks with preceding flow and masks.

Table 2: *Comparison with previous state-of-the-art methods on future image segmentation.* Our results consistently outperform the second-best results (across all methods) by +1.3% on EPIC-Kitchens and +0.7% on EGTEA in average F1 score. *: We re-implement the model to take raw video frames as inputs as our method (See more discussion in Sec. 4.3). The best results are highlighted with **boldface**, and the second-best results are underlined.

(a) Experimental Results on EPIC-Kitchens Dataset

Method	Epic-Kitchens (Precision/ Recall/ F1 Score)		
	short-term	middle-term	long-term
X2X [35]	68.69/ 69.35/ 69.02	40.81/ 50.61/ 45.18	28.14/ 45.76/ 34.85
ConvLSTM [46]	69.02/ 69.44/ 69.22	42.72 /51.78/ 46.82	30.01/ 48.01/ 36.94
FlowTrans [22]	69.38/ 69.70 /69.54	42.90/ 52.02/ 47.02	30.19/ 47.56/ 36.94
EgoGAN (Ours)	**70.89**/ **71.24**/ **71.07**	**43.79**/ **53.23**/ **48.05**	**31.39**/ **48.57**/ **38.14**

(b) Experimental Results on EGTEA Gaze+ Dataset

Method	EGTEA (Precision/ Recall/ F1 Score)		
	short-term	middle-term	long-term
X2X [35]	42.96/ 59.32/ 49.84	38.70/ 59.89/ 47.01	36.55/ 59.67/ 45.33
ConvLSTM [46]	44.55/ 59.43/ 50.93	38.28/ **63.54**/ 47.78	36.58/ 62.04/ 46.03
FlowTrans [22]	44.22/ 61.36/ 51.40	40.38/ 58.62/ 47.82	35.04/ **64.34**/ 45.37
EgoGAN (Ours)	**44.91**/ **61.48**/ **51.91**	**41.10**/ 59.90/ **48.75**	**38.16**/ 59.88/ **46.61**

•ConvLSTM [46] uses a Convolutional LSTM to model the temporal relationships of image features, and uses both the sequence of image features and the output of the ConvLSTM module for future image segmentation.

It is worth noting that the baseline methods [35,22,46] adopt a weaker backbone network than ours. To show that the performance gain of our method does not come from a stronger video feature encoder, we re-implement the above methods with the same I3D-Res50 backbone network as ours. Moreover, both FlowTrans and ConvLSTM assume accurate semantic segmentation of observable video frames is available as input, but our model seeks to forecast future hand segmentation using only raw video frames, and thus is a more challenging and practical setting. In addition, accurate semantic segmentation results on egocentric video frames are difficult to obtain due to the domain gap and lack of training data. Therefore, for a fair comparison, we implement the ConvLSTM and FlowTrans models to take the same input as our method. In our supplementary materials, we show that using the segmentation results from the pre-trained segmentation network as inputs will compromise the performance of FlowTrans and ConvLSTM.

The experimental results are summarized in Table 2. Among all baseline methods, FlowTrans achieves the best performance for short-term anticipation. However, it is less effective for long-term anticipation, due to the error accumulation of predicted future optical flow. ConvLSTM can better capture the long-term temporal relationship and thereby achieving the best baseline perfor-

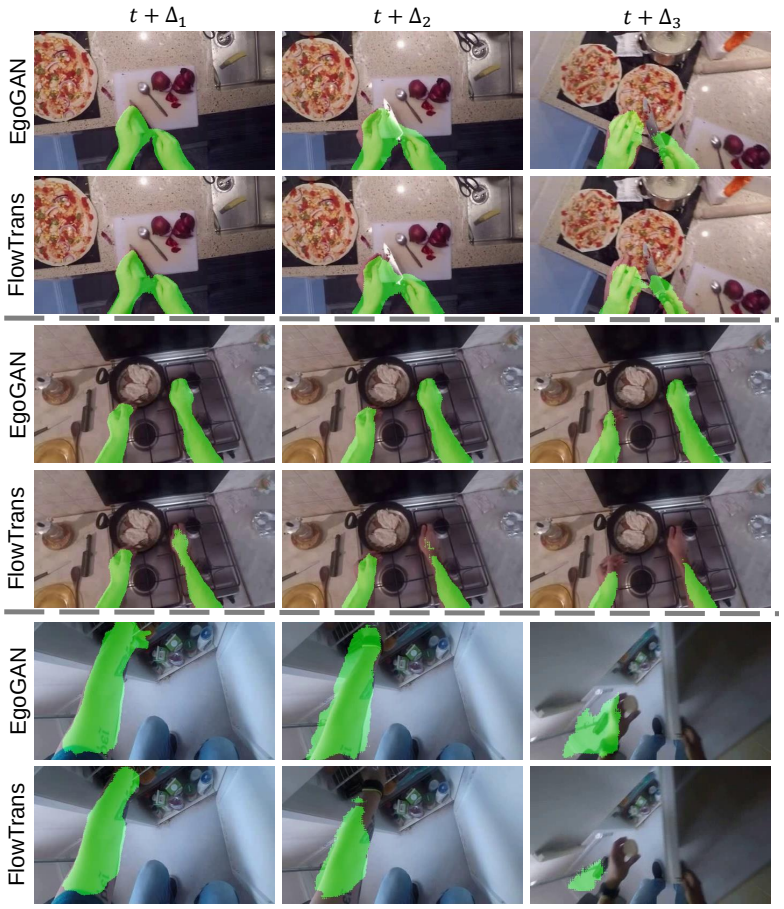

Fig. 3: *Visualization of our results.* From left to right, each column presents the future hand segmentation results of short-term ($t + \Delta_1$), middle-term ($t + \Delta_2$), and long-term($t + \Delta_3$) time steps from the EPIC-Kitchens dataset. Predictions from our method *EgoGAN* and the best baseline *FlowTrans* are presented in each sample. (See more discussion in Sec. 4.4)

mance for long-term anticipation. Instead of encoding the temporal connection with recursive prediction, we found that the 3D deconvolution operation is effective for capturing the temporal correlation of anticipation video segments, and in doing so, it helps capture the future hand masks in one shot. More importantly, our method outperforms previous best results (underlined in Table 2) by +1.5%/1.0%/1.2% and +0.5%/0.9%/0.6% in F1 score for short/middle/long term hand mask anticipation on EPIC-Kitchens and EGTEA, respectively. Once again, these results demonstrate the benefits of explicitly modeling future head motion with a GAN.

4.4 Discussion

Visualization. We visualize the results from both our method *EgoGAN* and the best baseline *FlowTrans* on EPIC-Kitchens in Fig. 3. Even though our proposed problem of future hand segmentation from egocentric video poses a formidable challenge in computer vision, our method can more accurately predict the hand region of future frames compared to FlowTrans, together with capturing the hand shape and poses. Notably, as the uncertainty increases with the anticipation time, our model may produce blurry predictions, yet can still robustly localize the hand region. The video demo in our supplementary material also suggests that our approach can produce satisfying results even when there are drastic hand and head movements. We conjecture that our model can better forecast the scene context change driven by head motion, and thereby more accurately predicts future hand masks.

Remarks. To summarize, our quantitative results indicate that future head motion carries important information for future hand movements. We show that explicitly modeling the underlying distribution of possible future hand movements with a GAN enables the model to predict the future hand masks more accurately. Another important takeaway is that our method is more effective than directly using a GAN for predicting future hand masks, as reported in Table 1. Furthermore, our visualizations demonstrate that our method can effectively predict future hand masks.

Limitations and Future Work. We also point out the limitations of our method. Since the hand mask ground truth does not differentiate left and right hands, our method cannot make separate predictions for the left and right hands. Recent work [48] does have the capability of localizing left and right hand bounding boxes separately during human-object interaction, and we plan to explore this direction in our future work on visual anticipation. In addition, our work does not explicitly exploit the action and object features for future hand prediction and will leave this for our future efforts. Nonetheless, our work investigates a novel and important problem in egocentric vision, and offers insight into visual anticipation and video pixel generation.

5 Conclusion

In this paper, we introduce the novel task of predicting a time series of future hand masks from egocentric videos. We present a novel deep generative model EgoGAN to address our proposed problem. The key innovation of our method is to use a GAN module that explicitly models the underlying distribution of possible future head motion for a more accurate prediction of future hand masks. We demonstrate the benefits of our method on two egocentric benchmark datasets, EGTEA Gaze+ and EPIC-Kitchens 55. We believe our work provides an essential step for visual anticipation as well as video pixel generation, and points to new research directions in the egocentric video.

Acknowledgments. Portions of this project were supported in part by a gift from Facebook. We thank Fiona Ryan for the valuable feedback.

References

1. Cai, M., Lu, F., Sato, Y.: Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In: CVPR (2020) 3, 9
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017) 8
3. Chandra, S., Couprie, C., Kokkinos, I.: Deep spatio-temporal random fields for efficient video segmentation. In: CVPR (2018) 3, 11
4. Chiu, H.k., Adeli, E., Niebles, J.C.: Segmenting the future. ICRA-L (2020) 4, 9
5. 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 dataset. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 753–771. Springer International Publishing, Cham (2018) 2, 3, 8
6. Dessalene, E., Devaraj, C., Maynord, M., Fermuller, C., Aloimonos, Y.: Forecasting action through contact representations from first person video. TPAMI (2021) 3
7. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: CVPR (2012) 3
8. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: ICCV (2011) 3
9. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 314–327. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 3
10. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: ICCV (2015) 4
11. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In: ICCV (2019) 1, 3, 4, 8
12. Gao, J., Yang, Z., Nevatia, R.: Red: Reinforced encoder-decoder networks for action anticipation. In: BMVC (2017) 4
13. Girdhar, R., Grauman, K.: Anticipative Video Transformer. In: ICCV (2021) 1
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS (2014) 4, 7
15. Gregor, K., Danihelka, I., Graves, A., Rezende, D., Wierstra, D.: Draw: A recurrent neural network for image generation. In: International Conference on Machine Learning. pp. 1462–1471. PMLR (2015) 4
16. Guan, J., Yuan, Y., Kitani, K.M., Rhinehart, N.: Generative hybrid representations for activity forecasting with no-regret learning. In: CVPR (2020) 4
17. Gui, L.Y., Wang, Y.X., Liang, X., Moura, J.M.F.: Adversarial geometry-aware human motion prediction. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 823–842. Springer International Publishing, Cham (2018) 4
18. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: CVPR (2018) 4
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 5
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015) 8

21. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017) 4, 6
22. 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) 4, 9, 11, 12
23. Kataoka, H., Miyashita, Y., Hayashi, M., Iwata, K., Satoh, Y.: Recognition of transitional action for short-term action prediction using discriminative temporal cnn feature. In: BMVC (2016) 4
24. Ke, Q., Fritz, M., Schiele, B.: Time-conditioned action anticipation in one shot. In: CVPR (2019) 1, 3, 4
25. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 201–214. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 4
26. Li, Y.: Learning embodied models of actions from first person video. Ph.D. thesis, Georgia Institute of Technology (2017) 9
27. Li, Y., Fathi, A., Rehg, J.M.: Learning to predict gaze in egocentric video. In: ICCV (2013) 6, 11
28. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 639–655. Springer International Publishing, Cham (2018) 2, 3, 9, 11
29. Li, Y., Liu, M., Rehg, J.M.: In the eye of the beholder: Gaze and actions in first person video. TPAMI (2021) 3, 8
30. Li, Y., Ye, Z., Rehg, J.M.: Delving into egocentric actions. In: CVPR (2015) 3
31. Liu, M., Ma, L., Somasundaram, K., Li, Y., Grauman, K., Rehg, J.M., Li, C.: Egocentric activity recognition and localization on a 3d map. arXiv preprint arXiv:2105.09544 (2021) 3
32. Liu, M., Tang, S., Li, Y., Rehg, J.M.: Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 704–721. Springer International Publishing, Cham (2020) 1, 3, 4, 8, 9
33. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015) 6
34. Luc, P., Couprie, C., LeCun, Y., Verbeek, J.: Predicting future instance segmentation by forecasting convolutional features. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 593–608. Springer International Publishing, Cham (2018) 4
35. Luc, P., Neverova, N., Couprie, C., Verbeek, J., LeCun, Y.: Predicting deeper into the future of semantic segmentation. In: ICCV (2017) 3, 5, 11, 12
36. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: CVPR (2016) 3
37. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014) 4
38. Moltisanti, D., Wray, M., Mayol-Cuevas, W., Damen, D.: Trespassing the boundaries: Labeling temporal bounds for object interactions in egocentric video. In: ICCV (2017) 3
39. Nilsson, D., Sminchisescu, C.: Semantic video segmentation by gated recurrent flow propagation. In: CVPR (2018) 3, 11
40. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: ICML (2017) 4

41. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016) 7
42. Pelz, J., Hayhoe, M., Loeber, R.: The coordination of eye, head, and hand movements in a natural task. Experimental brain research **139**(3), 266–277 (2001) 2
43. Pérez, J.S., Meinhardt-Llopis, E., Facciolo, G.: TV-L1 optical flow estimation. IPOL (2013) 8
44. Poleg, Y., Arora, C., Peleg, S.: Temporal segmentation of egocentric videos. In: CVPR (2014) 3
45. Poleg, Y., Ephrat, A., Peleg, S., Arora, C.: Compact CNN for indexing egocentric videos. In: WACV (2016) 3
46. Rochan, M., et al.: Future semantic segmentation with convolutional lstm. In: BMVC (2018) 4, 5, 12
47. Rodriguez, C., Fernando, B., Li, H.: Action anticipation by predicting future dynamic images. In: Leal-Taixé, L., Roth, S. (eds.) Computer Vision – ECCV 2018 Workshops. pp. 89–105. Springer International Publishing, Cham (2019) 4
48. Shan, D., Geng, J., Shu, M., Fouhey, D.: Understanding human hands in contact at internet scale. In: CVPR (2020) 14
49. Shen, Y., Ni, B., Li, Z., Zhuang, N.: Egocentric activity prediction via event modulated attention. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 202–217. Springer International Publishing, Cham (2018) 1, 3
50. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS (2014) 9
51. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) 5
52. Soo Park, H., Shi, J.: Social saliency prediction. In: CVPR (2015) 3
53. Soran, B., Farhadi, A., Shapiro, L.: Generating notifications for missing actions: Don't forget to turn the lights off! In: ICCV (2015) 3, 4
54. Tsai, Y.H., Yang, M.H., Black, M.J.: Video segmentation via object flow. In: CVPR (2016) 3, 11
55. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: CVPR (2018) 4
56. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: CVPR (2016) 4
57. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NeurIPS (2016) 6
58. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from static images using variational autoencoders. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 835–851. Springer International Publishing, Cham (2016) 4
59. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: ICCV (2017) 4, 6
60. Wang, W., Zhou, T., Porikli, F., Crandall, D., Van Gool, L.: A survey on deep learning technique for video segmentation. arXiv preprint arXiv:2107.01153 (2021) 3
61. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018) 5, 8
62. Xu, Y.S., Fu, T.J., Yang, H.K., Lee, C.Y.: Dynamic video segmentation network. In: CVPR (2018) 3, 11
63. Yagi, T., Mangalam, K., Yonetani, R., Sato, Y.: Future person localization in first-person videos. In: CVPR (2018) 3

64. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019) 3, 11
65. Yonetani, R., Kitani, K.M., Sato, Y.: Recognizing micro-actions and reactions from paired egocentric videos. In: CVPR (2016) 3
66. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017) 4
67. Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., Feng, J.: Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In: CVPR (2017) 2, 4, 6
68. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: CVPR (2021) 4
69. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: CVPR (2020) 4