Deep Future Gaze: Gaze Anticipation on Egocentric Videos Using Adversarial Networks

Mengmi Zhang[1,2], Keng Teck Ma[2], Joo Hwee Lim[2], Qi Zhao[3,1], and Jiashi Feng[1]

mengmi@u.nus.edu, {makt,joohwee}@i2r.a-star.edu.sg, qzhao@cs.umn.edu, elefjia@nus.edu.sg
[1]National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]University of Minnesota, USA

Abstract

We introduce a new problem of gaze anticipation on egocentric videos. This substantially extends the conventional gaze prediction problem to future frames by no longer confining it on the current frame. To solve this problem, we propose a new generative adversarial neural network based model, Deep Future Gaze (DFG). DFG generates multiple future frames conditioned on the single current frame and anticipates corresponding future gazes in next few seconds. It consists of two networks: generator and discriminator. The generator uses a two-stream spatial temporal convolution architecture (3D-CNN) explicitly untangling the foreground and the background to generate future frames. It then attaches another 3D-CNN for gaze anticipation based on these synthetic frames. The discriminator plays against the generator by differentiating the synthetic frames of the generator from the real frames. Through competition with discriminator, the generator progressively improves quality of the future frames and thus anticipates future gaze better. Experimental results on the publicly available egocentric datasets show that DFG significantly outperforms all well-established baselines. Moreover, we demonstrate that DFG achieves better performance of gaze prediction on current frames than state-of-the-art methods. This is due to benefiting from learning motion discriminative representations in frame generation. We further contribute a new egocentric dataset (OST) in the object search task. DFG also achieves the best performance for this challenging dataset.

1. Introduction

Egocentric video analysis [2], *i.e.* analyzing videos captured from the first person perspective, is an emerging field in computer vision which can benefit many applications, such as virtual reality (VR) and augmented reality (AR).

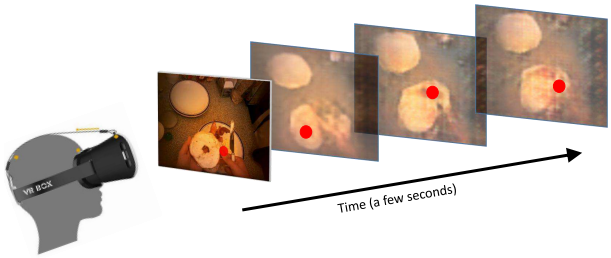


Figure 1. Problem illustration: gaze anticipation on future frames within a few seconds on egocentric videos. Given the current frame, the task is to predict the future gaze locations. Our proposed DFG method solves this problem through synthesizing future frames (transparent ones) and predicting corresponding future gaze locations (red circles).

One of the key components in egocentric video analysis is gaze prediction — the process of predicting the point of gaze (where human is fixating) in the head-centered coordinate system. Extending the gaze prediction problem to go beyond the current frame [37, 25], our paper presents the new and important problem of *gaze anticipation*: the prediction of gaze in future[1] frames of egocentric videos and proposes a promising solution.

Gaze anticipation enables the predictive computation and is useful in many applications, such as human-machine interaction [39, 9, 32], attention-driver user interface [22] and interactive advertisements [29]. For example, VR headsets, as one category of egocentric devices, require high computation power and fast speed for synthesizing virtual realities upon interaction from users [28, 8]. As gaze information reflects human intent and goal inferences [16, 38], gaze anticipation facilitates the computation-demanding systems to plan ahead on VR rendering with increased buffer time [39]. Thus, pre-rendering of the virtual scenes

[1]By "future" we mean within a few seconds.

based on anticipated gaze locations within the next few seconds provides smoother transitions in virtual reality and hence better user experience [28, 8].

We tackle gaze anticipation problem in two steps. Given the current frame, our proposed model, Deep Future Gaze (DFG), first generates future frames and then predicts the gaze locations on these frames.

As the dense optical flow between adjacent frames does not prorogate well to subsequent frames [27], we propose a generative adversarial network (GAN) based model for future frame generation through a competition between a generator (**GN**) and a discriminator (**D**) [27, 36]. Future frame generation of egocentric videos is a challenging task. Compared with the third-person videos where the background is often static, egocentric videos also involve complex background motion due to the head movements. We use a two-stream spatial-temporal convolution architecture (3D-CNN) in **GN** to explicitly untangle the foreground and the background motion. In the video generation approach [27], the aim is to generate "real" videos with the random noise as the input. Different from them, we have an additional constraint that our generated frames have to be based upon the input frame. Thus, we attach a 2D convolutional network (2D-CNN) before the **GN** to extract the latent representation of the current frame such that the motion dynamics of the generated frames is consistent with the current frame across time. Egocentric vision in a natural environment is a complex coordination process among head motion, gaze prediction, and body poses [23]. DFG learns automatically to model this coordination in **GN** without explicitly defining egocentric cues, such as hands, objects and task information. The rich information including egocentric motion dynamics on the generated future frames can then be useful for gaze anticipation.

D plays against **GN** by differentiating the synthetic frames of **GN** from the real frames. Through competition with **D**, **GN** progressively improves quality of the future frames based on the feedbacks from **D** and thus anticipates future gazes better.

Evaluations of DFG on public egocentric datasets show that DFG boosts up the performance of gaze anticipation to a significant extent surpassing all the well-established baselines. Additionally, DFG demonstrates its capacity of generalizing to the object search task on our new egocentric dataset (OST). OST is one of the largest egocentric datasets in the object search task with eyetracking information available to our best knowledge.

In summary, our paper has the following contributions:

- We introduce a new and important problem of gaze anticipation on egocentric videos.
- In order to tackle this new problem, we propose a new GAN-based model. A novel two-stream 3D-CNN is developed to explicitly untangle foreground and background motions in egocentric videos.
- Instead of handcrafting egocentric cues, such as hands and objects, our model automatically learns these cues during end-to-end training.
- We provide a new egocentric dataset downloadable at our website[2]. It is one of the largest in the object search task with eye-tracking information available.

2. Related Work

In this section, we review important works related to computational models of visual attention and gaze prediction on egocentric videos in particular.

2.1. Saliency Prediction

Computational saliency models are based on feature-integration theory [35] where low-level features, such as color, contrast and intensity, are combined. The first models were developed by Koch *et al*. [21] and Itti *et al*. [19]. Subsequent works [13, 41, 7, 14] further improve saliency map predictions via various methods such as graph-based saliency model [13] and boolean map based saliency [40]. The most recent saliency models leverage rich pools of semantic regions or objects in the scene from deep convolutional neural network [17, 26], whereas they focus on saliency prediction on static images and the motion information across frames has been discarded.

There are a few works exploiting top-down mechanisms. In [33], the contextual information from the scene was integrated with low-level features for saliency prediction. Borji *et al*. [3] explored a direct mapping from motor actions and low-level features to fixation locations in the driving simulation scenario where motor actions are from the top-down stream. In these cases, additional information other than egocentric videos is required.

2.2. Gaze Prediction on Egocentric Videos

Ba *et al*. [1] proposed to analyze human visual attention by exploring correlations between head orientation and gaze direction. Similarly, Yamada *et al*. [37] presented gaze prediction models and explored the correlation between gaze and head motion with the aid of external motion sensors. However, motion sensors may increase the loads and power consumption of wearable devices. The most recent model on gaze prediction in hand-object manipulation tasks was proposed by Yin *et al*. [25]. Hand detection and pose recognition provide primary egocentric cues in their model. Since egocentric cues are predefined, their model may not generalize well to various egocentric activities especially when hands are not involved.

[2] https://github.com/Mengmi/deepfuturegaze_gan

To the best of our knowledge, we are the first to tackle gaze anticipation problem on egocentric videos. We propose a novel GAN-based model which can learn essential egocentric cues automatically during the training phase. Different from third person videos, head motion results in moving background in egocentric videos. Thus, we adapt the two-stream video model [31, 36] with both streams replaced by 3D-CNN to explicitly untangle foreground and background motions.

3. Our Model

In this section, we first introduce an overview of our proposed model, Deep Future Gaze (DFG), and then give the detailed analysis of its architecture. We provide the training and implementation details in the end.

3.1. Architecture Overview

Given the current frame as the input, we aim to output a sequence of anticipated gaze locations in the next few seconds. To address this challenging problem, we propose a generative adversarial networks (GAN) [27, 36] based model, to generate future frames and then to predict their corresponding temporal saliency maps, *i.e.*, spatial probabilistic maps of gaze locations across time where the spatial coordinates with the maximum probability are output as the anticipated gaze locations. DFG consists of two networks: the Generator Network (**GN**) and the Discriminator Network (**D**) as shown in Figure 2. In **GN**, there are two modules: **Future Frame Generation Module (G)** and **Temporal Saliency Prediction Module (GP)**. See Supplementary Material for architecture details.

3.2. The Generator Network (GN)

The goal of **GN** is to produce a sequence of N subsequent frames $I_{t+1,t+N}$ from a latent representation $h(I_t)$ of the current frame I_t in **G** and N temporal saliency maps $S_{t+1,t+N}$ from $I_{t+1,t+N}$ in **GP**. Here the latent representation $h(I_t)$ is learned from a 2D-CNN. In order to identify the foreground motions (hands and objects) out of the complex background motion due to the head movements, we propose a two-stream generator architecture. To avoid the error in the frame generation accumulating from one frame to another, **G** is designed to generate a sequence of N future frames at once instead of a system where the generated frame I_{t+1} is fed back as the input to generate the subsequent frame I_{t+2}. The number of predicted frames N is application dependent. We select 32 frames or about 2.5 seconds as we believe such duration is adequate for practical applications. The complete analysis regarding the performance of our model versus number of output frames is presented in Section 4.6.

We use 3D-CNN in two streams for learning motion representations. Meanwhile, fractionally strided convolution layers (upsampling layers) are added after the convolution to preserve proper spatial and temporal resolution for the output frame sequence. The equation for generating the sequence of N predicted frames $I_{t+1,t+N}$ is

$$\begin{aligned} I_{t+1,t+N} =& F(h(I_t)) \odot M(h(I_t)) \\ &+ (1 - M(h(I_t))) \odot B(h(I_t)), \end{aligned} \tag{1}$$

where $\odot$ is the elementwise-multiplication operation, $F(\cdot)$ represents the foreground generation model and $B(\cdot)$ represents the background generation model. $M(\cdot)$ is a spatial-temporal mask untangling foreground and background motion where its pixel value ranges from $[0, 1]$. In particular, 1 indicates foreground and 0 indicates background. Both $F(\cdot)$ and $B(\cdot)$ generate a sequence of N predicted RGB-colored frames, each frame with dimension $3 \times W \times H$ where W and H are the width and the height of the predicted frame respectively. Foregrounds and backgrounds of predicted frames get merged by masks $M(\cdot)$ of dimension $N \times 1 \times W \times H$ replicated across 3 color channels to produce $I_{t+1,t+N}$. The foreground, background and mask models are parameterized by 3D-CNN. The foreground model and the mask model share the same weights until the last layer which has two branches, one for foreground generation for N frames with 3 color channels and one for the mask generation for N frames with single channels. The background generation model employs another separate 3D-CNN.

As the rich information including the learnt egocentric motion dynamics on the generated future frames is useful for visual attention in egocentric videos, we adopt these features for gaze anticipation. Thus, **G** is followed by **GP** to generate temporal saliency maps of dimension $N \times 1 \times W \times H$.

3.3. The Discriminator Network (D)

Generating N frames implies the need of a large number of pixels. This is an extremely difficult task when only a single frame is given. To enhance the quality of generated frames, DFG employs **D** as a competitor to **G**, by providing the additional feedbacks to **G** [27, 36].

D aims to distinguish the synthetic examples from the real ones. There are two criteria for the synthetic frames to be "real": first, the semantics from the scene are coherent across space (e.g. no table surface inside the refrigerator); second, the motions from both the foreground and the background are consistent across time (e.g. hand movements have to be smooth). Thus, **D** follows the same architecture as the foreground generation model other than replacing all the upsampling layers with the convolution layers as detailed in Supplementary Material and this architecture has also been shown to be effective in [36]. The output is a binary label indicating whether the input frame is fake or real.

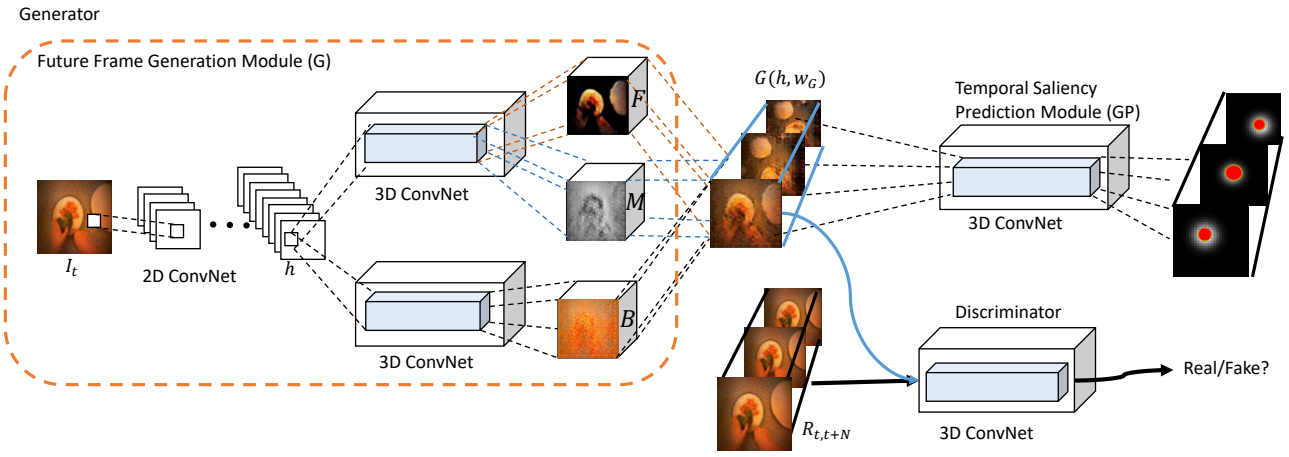


Figure 2. Architecture of our proposed Deep Future Gaze (DFG) model. It contains **Generator Network** and **Discriminator Network**. In **Future Frame Generation Module** of **Generator Network**, latent representation of the current frame I_t is extracted by 2D ConvNet. To explicitly untangle foreground and background, it then branches into two streams: one for learning the representation for the foreground and the mask; one for learning the representation of the background. These 3 streams are combined to generate future frames (blue boundaries). Based on the generated frames, **Temporal Saliency Prediction Module** predicts the anticipated gaze location (red dots). As a competitor to the generator, **Discriminator Network** uses a 3D ConvNet to distinguish the generated frames from real frames $R_{t,t+N}$ (black boundaries) by classifying its inputs to real or fake.

3.4. Training and Implementation Details

Training We train DFG end-to-end by stochastic gradient descent with learning rate 0.00005 and momentum 0.5. Adam Optimizer [20] is used. **G** and **D** play against each other. **G** is designed to predict future frames as "real" as possible to fool **D**, while **D** strives to tell real frames from the generated ones. These two networks try to minimize the maximum payoff of its opponent with respect to their network parameters w_D and w_G respectively. In addition, we add another *L1* loss term to ensure that the first generated video frame is visually consistent with the input frame without the over-smoothing artifacts. A hyper-parameter λ is used for tuning the weight of losses between the min-max game and the consistency term. Both networks are trained alternatively. The objective function for **D** is

$$\begin{aligned}\min_{w_D} f_D(R_{t:t+N}, h) \triangleq\ & L_{ce}(D(R_{t:t+N}; w_D), 1) \\ & + L_{ce}(D(G(h; w_G)), 0),\end{aligned} \tag{2}$$

where h denotes the hidden representation $h(I_t)$ of input frame I_t, $R_{t:t+N}$ represents the real frames and the binary cross entropy loss L_{ce} is defined as

$$L_{ce}(\hat{Y}, Y) = Y \log(\hat{Y}) + (1 - Y) \log(1 - \hat{Y}), \tag{3}$$

where $Y \in \{0, 1\}$ denotes real or fake and $\hat{Y} \in [0, 1]$ denotes the output from **D**.

As the opponent of **D**, **G** needs to satisfy two requirements: 1) the generated outputs should be real enough to fool **D**; 2) the initial output of the generated frames should be visually consistent with the current frame. The objective function for training **G** is thus formulated as

$$\begin{aligned}\min_{w_G} f_G(I_t) \triangleq\ & L_{ce}(D(G(h; w_G)), 1) \\ & + \lambda \| I_t - G(I_t; w_G) \|_1,\end{aligned} \tag{4}$$

where λ is set as 0.1 which shows to achieve the best performance in our case. $\| \cdot \|_1$ denoting L1 distance is preferred over the mean square error which results in over-smoothing in the frame generation [27].

Meanwhile, **GP** takes $I_{t+1,t+N}$ as input to generate temporal saliency maps. **GP** is trained in a supervised approach using Kullback-Leibler divergence (KLD) loss function:

$$KLD(P_i, Q_i) = \sum_x \sum_y P_i(x, y) \log \left[\frac{P_i(x, y)}{Q_i(x, y)} \right], \tag{5}$$

where P_i is the temporal fixation map and Q_i is the temporal saliency map for the $(t + i)$th frame.

Implementation Details DFG is developed based on [36] in Torch. The source code is available at our website[2]. We train everything from zero with the input frame size being $3 \times 64 \times 64$. The batch size is 32. The latent representation $h(I_t)$ is of dimension $1024 \times 4 \times 4$ after 5 layers of 2D convolution layers for encoding image representation. We normalize all videos to be within the range $[-1, 1]$.

Gaze prediction on current frame DFG can also be used for gaze prediction on the current frame. Since **G** outputs a sequence of generated frames where the first frame must be consistent with the input frame due to *L1* distance loss in

Equation(4), we take the spatial coordinate with the maximum probability in the first predicted temporal saliency map as the predicted gaze location on the current frame.

4. Experiments

We test DFG on gaze anticipation as well as gaze prediction over current frames on public datasets using standard evaluation metrics. To explore whether DFG can be generalized well for other tasks in egocentric contexts, we contribute another dataset (OST) in the object search task. We provide detailed analysis of DFG through ablation study and visualization of the learnt convolution filters. In the end, we demonstrate our anticipated gazes are useful in egocentric activity recognition.

4.1. Datasets

GTEA Dataset [11] This dataset contains 17 sequences on meal preparation tasks performed by 14 subjects. Each video clip lasts for about 4 minutes with the frame rate 15 fps and frame resolution 480×640. The subject is asked to prepare meals freely. Same as Yin *et al*. [25], we use videos 1, 4, 6-22 as training set and the rest as test set.

GTEAplus Dataset [25] This dataset consists of 7 meal preparation activities. There are 5 subjects, each performing these 7 activities. Each video clip takes 10 to 15 minutes on average with frame rate 12 fps and frame resolution 960×1280. We do 5-fold cross validation across all 5 subjects and take their average for evaluation as [25].

Our Dataset in Object Search Tasks (OST) Due to lack of egocentric datasets with gaze tracking enabled, we contribute this new dataset for the object search task. This dataset consists of 57 sequences on search and retrieval tasks performed by 55 subjects. Each video clip lasts for around 15 minutes with the frame rate 10 fps and frame resolution 480×640. Each subject is asked to search for a list of 22 items (including lanyard, laptop, *etc.*) and move them to the packing location (dining table). Details about the 22 items are provided in Supplementary Material. See Figure 3 for exemplar frames. We select frames near the packing location and use those from videos 1 to 7 as test set and the rest for training and validation.

To the best of our knowledge, this is one of the largest egocentric datasets on the object search task with eye-tracking information available. Compared with GTEA and GTEAplus, our dataset involves larger head motions and the human subjects have to walk and look for objects in the search list with hands appearing less frequently.

4.2. Evaluation Metrics

We use two standard evaluation metrics on gaze anticipation in egocentric videos: Area Under the Curve (AUC) [4] and Average Angular Error (AAE) [30] defined as below:

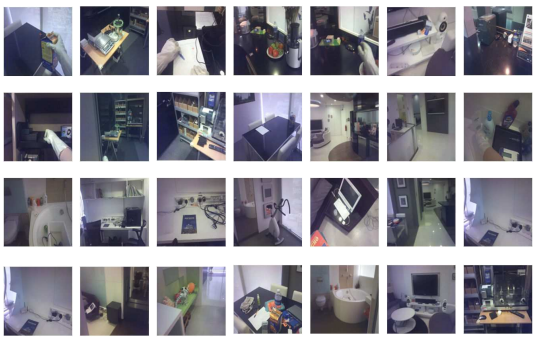


Figure 3. Sample frames from our introduced object search dataset (OST). It covers various rooms in a fully furnished 2-bedroom apartment and includes 22 searching items. For our experiments, only frames near the packing location (dining table) are selected.

Area Under the Curve (AUC) is the most commonly used saliency evaluation metric. It measures the area under a curve of true positive versus false positive rates under various threshold values on saliency maps.

Average Angular Error (AAE) is the angular distance between the predicted gaze location and the ground truth.

4.3. Baselines

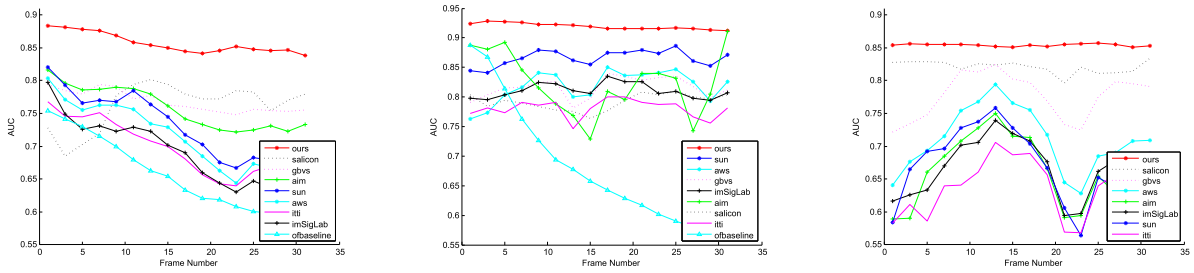
We create several competitive baselines as follows:

First, to show the effectiveness of end-to-end learning where all the parameters are trained jointly, we use **G** to generate future frames after the training phase and compare DFG with state-of-the-art saliency prediction algorithms on these frames including Graph-based Visual Saliency (GBVS) [13], Natural Statistics Saliency (SUN) [41], Adaptive Whitening Saliency (AWS) [12], Attention-based Information Maximization (AIM) [6], Itti's Model (Itti) [18], and Image Signature Saliency (ImSig) [15].

Second, SALICON [17] is a deep learning architecture for saliency prediction on static images. We train SALICON from scratch on the egocentric datasets by using real frames and their corresponding fixation maps. After that, the pre-trained SALICON model is tested on our generated frames for gaze anticipation.

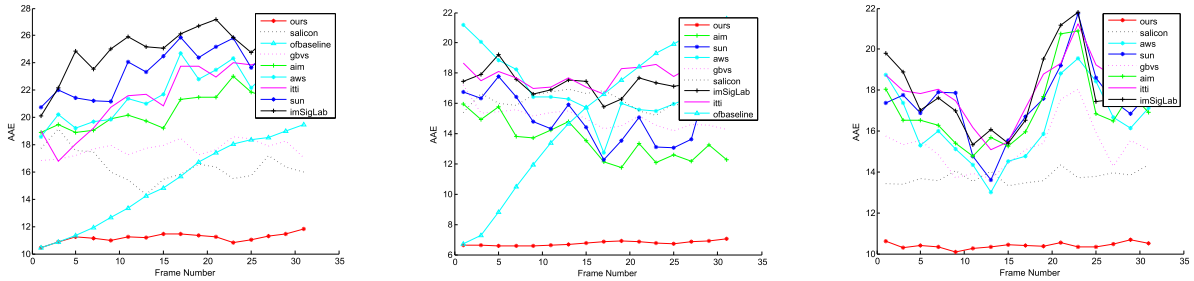
Third, we create another baseline (OpticalShift) to study the effect of temporal dynamics. We use our model to predict gaze on the current frame and compute the dense optical flow between the previous frame and the current frame using [5]. The predicted gaze is then warped to the future frames by shifting it based on the flow at that position as the future gaze locations.

Fourth, we include the graph-based method to model gaze transition dynamics as proposed by [25] for gaze prediction on current frames in GTEA and GTEAplus. We exclude this method on OST since the required hand annotations by [25] are not available. We also cannot extend this method to gaze anticipation problem.



(a) GTEA Dataset (2.1 sec ahead) (b) GTEAplus Dataset (2.7 sec ahead) (c) Our OST Dataset (3.2 sec ahead)

Figure 4. Evaluation of Gaze Anticipation using Area Under the Curve (AUC) on the current frame as well as 31 future frames in GTEA, GTEAplus and Our OST Dataset. Larger is better. The algorithms in the legend are introduced in Section 4.3.



(a) GTEA Dataset (2.1 sec ahead) (b) GTEAplus Dataset (2.7 sec ahead) (c) Our OST Dataset (3.2 sec ahead)

Figure 5. Evaluation of Gaze Anticipation using Average Angular Error (AAE) on the current frame as well as 31 future frames in GTEA, GTEAplus and Our OST Datasets. Smaller is better. The algorithms in the legend are introduced in Section 4.3.

4.4. Results on Gaze Anticipation

DFG surpasses all the competitive baselines significantly in gaze anticipation. We report the quantitative evaluation results in Figure 4 (AUC) and 5 (AAE). On all three datasets, DFG outperforms all the competitive baselines by 31%, 50% and 24% in relative advance (RA) in AAE and 21%, 5% and 3% in RA in AUC with respect to the best baseline (BB) as shown in Figure 4 and 5. RA in percentage is computed as

$$\mathrm{RA}(\mathrm{OUR}, \mathrm{BB}) = \frac{\|\sum_{i=1}^{N} \mathrm{OUR}_i - \sum_{i=1}^{N} \mathrm{BB}_i\|}{\sum_{i=1}^{N} \mathrm{BB}_i}, \quad (6)$$

where N=32 is the number of generated future frames, OUR_i is the metric score of our model and BB_i is the metric score of BB on the ith future frame.

Qualitative results in Figure 6 demonstrate that DFG learns to untangle foreground and background motions. In the foreground, both the hand and the object (the bun) get highlighted. As the high intensity value on the mask denotes the foreground, the manipulation point (the control point where the subject is manipulating the object with hands) shows the highest activation on the mask whereas the background (the table surface) is uniform over time as shown in the darker regions of the mask.

Though SALICON learns an abundance of semantic information, it excludes temporal dependencies which are crucial for gaze anticipation on egocentric videos. Although SALICON has performed better than conventional saliency prediction methods, its performance is inferior to DFG

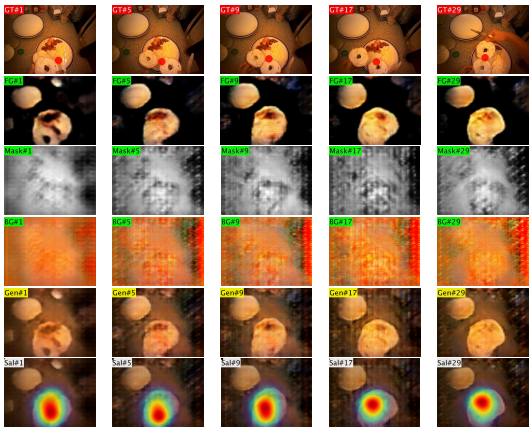


Figure 6. Example results of gaze anticipation on GTEAplus dataset. Our DFG model produces 31 future frames (2.67 seconds ahead) based on the current frame. Frames #1, 5, 9, 17, 29 are shown (left to right columns). The topmost row shows the ground truth with red circle denoting human gaze locations. Row 2, 3, 4 show the foreground $F(\cdot)$, the mask $M(\cdot)$, and the background $B(\cdot)$ learnt by **Generator Network** respectively. Row 5 shows the generated future frames. Row 6 shows the corresponding predicted temporal saliency maps. Best viewed in color.

which learns spatial-temporal information.

For OpticalShift, we observe that its AUC and AAE curves drop monotonically. It confirms that the optical flow computed from the current state cannot adapt to the complexity of the temporal dynamics in longer time periods.

We often observe a strong center bias (CB) in egocentric videos. This is due to the fact that egocentric videos are captured from the first person perspective. Humans always move their heads to attend to the regions of interest. In this case, gazes often align with head orientations. Thus, gaze shift in the large distance gets compensated by head movements with small gaze shifts. The statistics of amplitudes for head and gaze motions in our test sets for GTEA and GTEAplus datasets are provided in Supplementary Material. To validate DFG predicts more than CB, we fit a Gaussian mask in the center to generate temporal saliency maps. As AUC favors CB, we use sAUC to compare our model with CB and report sAUC scores as: DFG (0.58) and CB (0.5) in GTEA, DFG (0.61) and CB (0.5) in GTEAplus as well as DFG (0.56) and CB (0.49) in OST. It confirms that our model incorporates various egocentric information and motion dynamics for gaze anticipation rather than CB. Based on the statistics of the two motions in our testsets, it also shows that both **GP** and **G** are critical for better gaze anticipation by estimating the two motions separately.

4.5. Results on Current Frame Gaze Prediction

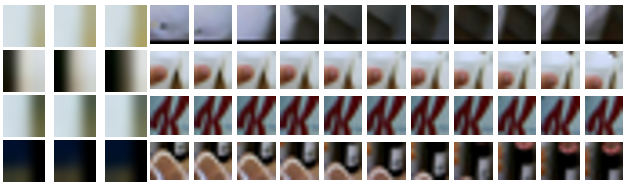
We compare DFG with state-of-the-art saliency prediction algorithms in Section 4.3 on real frames in the testsets of three datasets and report both AAE and AUC scores of gaze prediction on current frames in Table 1. Results show that DFG performs competitively well or better than the-state-of-the-arts even without explicitly specifying egocentric cues, such as hands and objects of interest. Moreover, different from the traditional methods, our model takes the current frame as the only input without the past information. We observe that AAE scores are lower than Yin *et al.* [25] on GTEA and this may be due to the small number of training samples compared with GTEAplus.

4.6. Ablation Study

In order to study the effect of the individual component of DFG, we do an ablation study and test on GTEA by removing *only* one component in DFG at one time while the rest of the architecture remains the same. There are three tests: (1) we replace the two-stream 3D-CNN in **G** with the same structure as [36], *i.e.* the background stream is "static" while the foreground stream remains the same; (2) we train **GP** directly on real frames and test **GP** on the generated frames from **G**; (3) we remove **D** and we only use L1 distance loss for future frame generation. Scores for gaze anticipation in AAE and AUC are averaged every 8 frames across 32 frames as shown in Table 2.

	GTEAplus		GTEA		Our OST	
Metrics	AUC	AAE	AUC	AAE	AUC	AAE
DFG(ours)	**0.952**	**6.6**	**0.883**	10.5	**0.854**	**10.6**
Yin [25]	0.867	7.9	0.878	**8.4**	-	-
SAL [17]	0.818	15.6	0.761	16.5	0.850	13.3
GBVS [13]	0.803	14.7	0.769	15.3	0.706	18.8
AWS [12]	0.824	14.8	0.775	17.5	0.563	22.8
AIM [6]	0.756	15.0	0.821	14.2	0.773	17.0
SUN [41]	0.842	14.7	0.802	18.1	0.527	25.0
Itti [18]	0.753	19.9	0.747	18.4	0.615	19.0
ImSig [15]	0.786	16.5	0.782	19.0	0.555	24.2

Table 1. Evaluation of gaze prediction on the current frame. We compare our model with state-of-the-arts using standard metrics Area Under the Curve (AUC) and Average Angular Error (AAE) on GTEA, GTEAplus and our OST dataset respectively. The algorithms listed are introduced in Section 4.3. (Number denoted in bold is the best.)



(a) GP1 (b) GP4

Figure 7. Visualization of the convolution filters in the first (GP1) and the second last (GP4) 3D convolution layers of **Temporal Saliency Prediction Module** in our DFG model. Subfigure 7a: the filters in the first 3D convolution layer show low-level features, such as edges. Subfigure 7b: the regions of salient objects are highly activated in the second last convolution layer, such as the fonts on the oatmeal box.

Angular Average Error (AAE)				
Frame#	# 1−8	# 9−16	# 17−24	# 25−32
Our Best	11.0	11.3	11.3	11.5
One-stream	12.3	11.9	12.2	12.1
Replace(GT)	12.8	13.3	13.9	13.9
Remove(D)	11.3	11.8	12.3	12.4
Area Under the Curve (AUC)				
Frame#	# 1−8	# 9−16	# 17−24	# 25−32
Our Best	0.88	0.86	0.85	0.84
One-stream	0.86	0.85	0.85	0.85
Replace(GT)	0.84	0.82	0.80	0.81
Remove(D)	0.86	0.84	0.82	0.81

Table 2. Ablation Study. From top to bottom, the evaluated models are: our best, one stream replaced by 2D convolution net; our model with **Future Frame Generation Module** removed and ground truth frames as direct inputs for training and future frames for testing; and the discriminator removed.

The first ablation study on changing the background stream to a static one leads to an increase of 0.7 in AAE. This implies the two-stream 3D-CNN in **G** is essential for learning foreground and background motions which can further improve gaze anticipation accuracy. The second ablated model with **GP** trained on real frames performs worse with an increase of 2 in AAE than DFG. In DFG, **GP** is attached after **G** for temporal saliency map prediction using end-to-end training. However, **GP** in the second ablated model trained only on real frames cannot perform well since it cannot learn the essential features on the generated

frames. It demonstrates that the features on the generated frames are different from those on real frames and hence, end-to-end training is necessary for **GP** to learn these essential features on the generated frames. The third ablation study with **D** removed shows an increases of 1 in AAE from the 17th frame onwards. This demonstrates that **D** is important as the feedback to **GN** which provides the additional constraints such that **G** can generate more "realistic" future frames in longer time duration. These "realistic" future frames are critical for gaze anticipation.

Moreover, to study the effectiveness of GAN-based architecture, we develop a few more comparative methods. See Supplementary Material for results and implementation details. First, we compare DFG with SalDirect: a 3D-ConvNet directly modeling gaze anticipation. Secondly, we develop a new model (SalFusion) which averages the temporal saliency maps from both SalDirect and DFG to generate the final temporal saliency maps. Results show DFG outperforms SalDirect in both AAE and AUC. It suggests GAN has essential contributions to gaze anticipation. SalFusion outperforms two composite models which confirms that the learnt motion cue from GANs is important and complementary to the cues learned directly from SalDirect.

It is also observed that the gaze movement on individual frames is dependent on their previous states; *e.g.* to anticipate gaze on the frame $t+32$, we need to consider gaze transitions across frames by also anticipating gaze on frames t to $t+31$. For verification, we created one baseline: train SALICON model, a 2D-ConvNet, directly for gaze anticipation at time $t+16$ and $t+32$ using their respective ground truth at time $t+16$ and $t+32$. See Supplementary Material for results. DFG performs much better than SALICON. This suggests the temporal dependence across frames plays fundamental roles in gaze anticipation in egocentric videos and future frame generation using GANs is useful.

In video analysis, the number of consecutive frames is a key parameter in practice. We study the effect of the number of frames on gaze anticipation. See Supplementary Material for implementation and result details. From the results, we observe that given an input frame, in order to anticipate gaze on subsequent L frames, models trained with $L+K$ frames will perform better as K increases. This is because **GP** can learn the temporal dynamics with more information flowing back from the future K frames. In the extreme scenario where $L=1$ and $K=0$, this architecture will be simplified as the feedforward 2D-CNN, similar as SALICON, and hence produces lower gaze anticipation performance.

4.7. Visualization

As **GP** estimates temporal saliency maps based on the generated frames, we analyze the learnt convolution filters in **GP** and align the observations with human bottom-up visual attention mechanism (VA). See Supplementary Materials for implementation details. We observe that the filters in the first convolution layer of **GP** learn the low level features, such as edges and regions of high contrast. This observation aligns well with VA which is driven by low level features at the initial stage according to [35]. More interestingly, we also find the learnt features change across time, e.g. the black region increases from left to right across time (row 2 in Figure 7a) and the brightness in the bottom regions decay across time (row 4 in Figure 7a). This demonstrates DFG learns motion dynamics such as translation and the gradient change of surfaces. As the level of convolution layers increases, we can see more complex patterns. In the second last layer, the regions containing semantic information get activated with some examples shown in Figure 7b. This includes salient objects, such as the white bowl, the tip of the milk box, the fonts on the oatmeal box and the bread with peanut butter. Overall, we infer that DFG not only learns egocentric cues in the spatial domain but also motion dynamics in the temporal domain.

4.8. Gaze-aided Egocentric Activity Recognition

Egocentric gaze can help first person's activity recognition. We adapt the C3D network [34] with integration of our anticipated gaze locations to egocentric activity recognition. Results show that our gaze-aided model (28.5%) significantly surpasses the initial C3D network [34] (26.9%), STIP [24] (14.9%), Cuboids [10] (22.7%), as well as one baseline (13.6%) on GTEAplus [3]. See Supplementary Material for implementation details and results analysis.

5. Conclusion

We present a new challenging gaze anticipation problem on future frames as an extension of the gaze prediction problem on current frames on egocentric videos. We develop DFG built upon GAN for solving this problem. We make great improvements on the existing works of GAN. To explicitly learn foreground and background motions in egocentric settings, we propose a two-stream 3D-CNN in the generator network. The qualitative results show that our model learns to untangle these motions. We evaluate our model using standard metrics and our performance surpasses all the competitive baselines significantly.

6. Acknowledgements

This work was supported by the Reverse Engineering Visual Intelligence for cognitive Enhancement (REVIVE) programme funded by the Joint Council Office of A*STAR, National University of Singapore startup grant R-263-000-C08-133 and Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112. We also like to thank Yin Li for his help in replicating the experimental setup in [25].

[3] $(\cdot)$ refers to the corresponding activity recognition rate

References

[1] S. O. Ba and J.-M. Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, 2011. 2

[2] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):744–760, 2015. 1

[3] A. Borji, D. N. Sihite, and L. Itti. Probabilistic learning of task-specific visual attention. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 470–477. IEEE, 2012. 2

[4] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *IEEE ICCV*, pages 921–928. IEEE, 2013. 5

[5] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In *CVPR*, pages 41–48. IEEE, 2009. 5

[6] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2005. 5, 7

[7] N. D. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3):5–5, 2009. 2

[8] D. Christopoulos, A. Gaitatzes, and G. Papaioannou. Image-based techniques for enhancing virtual reality environments. In *2nd International Workshop on ICT's, Arts and Cultural Heritage*, 2003. 1, 2

[9] W. Ding, P. Chen, H. Al-Mubaid, and M. Pomplun. A gaze-controlled interface to virtual reality applications for motor- and speech-impaired users. *HCI International, San Diego, CA*, 2009. 1

[10] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72. IEEE, 2005. 8

[11] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012. 5

[12] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, 2012. 5, 7

[13] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 2, 5, 7

[14] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *TPAMI*, 34(1):194–201, 2012. 2

[15] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):194–201, 2012. 5, 7

[16] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology*, 6, 2015. 1

[17] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE ICCV*, pages 262–270, 2015. 2, 5, 7

[18] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10):1489–1506, 2000. 5, 7

[19] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, (11):1254–1259, 1998. 2

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[21] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987. 2

[22] M. Kumar. *Gaze-enhanced user interface design*. PhD thesis, Citeseer, 2007. 1

[23] M. F. Land. The coordination of rotations of the eyes, head and trunk in saccadic turns produced in natural situations. *Experimental brain research*, 159(2):151–160, 2004. 2

[24] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005. 8

[25] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3216–3223, 2013. 1, 2, 5, 7, 8

[26] Y. Lin, S. Kong, D. Wang, and Y. Zhuang. Saliency detection within a deep convolutional architecture. 2

[27] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 2, 3, 4

[28] F. Multon, L. France, M.-P. Cani-Gascuel, and G. Debunne. Computer animation of human walking: a survey. *The journal of visualization and computer animation*, 10(1):39–54, 1999. 1, 2

[29] R. Ohme, M. Matukin, and B. Pacula-Lesniak. Biometric measures for interactive advertising research. *Journal of Interactive Advertising*, 11(2):60–72, 2011. 1

[30] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1153–1160, 2013. 5

[31] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014. 3

[32] M. Stengel, S. Grogorick, M. Eisemann, E. Eisemann, and M. A. Magnor. An affordable solution for binocular eye tracking and calibration in head-mounted displays. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 15–24. ACM, 2015. 1

[33] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006. 2

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *arXiv preprint arXiv:1412.0767*, 2014. 8

[35] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 2, 8

[36] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. *arXiv preprint arXiv:1609.02612*, 2016. 2, 3, 4, 7

[37] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki. Attention prediction in egocentric video using motion and visual saliency. In *Pacific-Rim Symposium on Image and Video Technology*, pages 277–288. Springer, 2011. 1, 2

[38] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in psychology*, 4:917, 2013. 1

[39] R. C. Zeleznik, A. S. Forsberg, and J. P. Schulze. Look-that-there: Exploiting gaze in virtual reality interactions. Technical report, Technical Report CS-05, 2005. 1

[40] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 153–160, 2013. 2

[41] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008. 2, 5, 7