Anticipating Where People will Look Using Adversarial Networks

Mengmi Zhang, *Student Member, IEEE*, Keng Teck Ma, *Member, IEEE*, Joo Hwee Lim, *Senior Member, IEEE*, Qi Zhao, *Member, IEEE*, and Jiashi Feng, *Member, IEEE*

Abstract—We introduce a new problem of gaze anticipation on future frames which extends the conventional gaze prediction problem to go beyond current frames. To solve this problem, we propose a new generative adversarial network based model, Deep Future Gaze (DFG), encompassing two pathways: DFG-P is to anticipate gaze prior maps conditioned on the input frame which provides task influences; DFG-G is to learn to model both semantic and motion information in future frame generation. DFG-P and DFG-G are then fused to anticipate future gazes. DFG-G consists of two networks: a generator and a discriminator. The generator uses a two-stream spatial-temporal convolution architecture (3D-CNN) for explicitly untangling the foreground and 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 distinguishing the synthetic frames of the generator from the real frames. Experimental results on the publicly available egocentric and third person video datasets show that DFG significantly outperforms all competitive baselines. We also demonstrate that DFG achieves better performance of gaze prediction on current frames in egocentric and third person videos than state-of-the-art methods.

Index Terms—Egocentric videos, gaze anticipation, generative adversarial network, saliency, visual attention

1 Introduction

EGOCENTRIC video analysis [1], 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). 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 [2], [3], our paper presents the new and important problem of *gaze anticipation* (Fig. 1): the prediction of gaze in future frames of egocentric videos within a few seconds and proposes a promising solution which is further developed from [4].

Gaze anticipation enables the predictive computation and is useful in many applications, such as human-machine interaction [5], attention-driver user interface [6] and interactive advertisements [7]. 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 [8]. Gaze anticipation facilitates the computation-demanding systems to plan ahead on VR rendering with increased buffer time [9]. Thus, pre-rendering of the virtual scenes based on anticipated gaze locations within the next few seconds provides smoother presentations in virtual reality and hence better user experience [8]. In interactive advertisement design [7], gaze anticipation could also assist remote information server in pre-fetching contextual e-advertisements and prompting to the consumers without noticeable time delays.

As gaze information reflects human intent and goal inferences [10], gaze anticipation also reduces users' reaction time with proactive feedbacks. It becomes critical especially in life-threatening scenarios, such as elderly fall prevention and collision avoidance in car driving. With gaze anticipation, the assistive system could anticipate the elderly's intention in navigation and alert them to be cautious about unnoticed hazards in front. Similarly, gaze anticipation could also be implemented in driver attention alert system and provide the proactive feedbacks to the drivers about unattended obstacles on the roads ahead.

Gaze, as a perceptual variable, cues attention. Attention can be categorized into two distinct functions: the bottom-up attentional guidance driven by external stimuli due to their inherent features relative to the backgrounds, such as the visual contrast; and the top down attention mechanism according to the current goals and purposeful plans, such as the navigation task towards the driver's desired destination location. Inspired by these attention mechanisms, we tackle gaze anticipation problem in two streams. Given the current frame, our proposed model, Deep Future Gaze (DFG), generates future frames using generative adversarial network

- M. Zhang is with the National University of Singapore, Singapore 119077, and also with the Institute for Infocomm Research (I^{2}R), A*STAR, Singapore 138632. E-mail: mengmi@u.nus.edu.
- K.T. Ma and J.H. Lim are with the A*AI and I^{2}R, A*STAR, Singapore 138632. E-mail: makt@scei.a-star.edu.sg, joohwee@i2r.a-star.edu.sg.
- Q. Zhao is with the University of Minnesota, Minneapolis, MN 55455, and also with the National University of Singapore, Singapore 119077. E-mail: qzhao@cs.umn.edu.
- J. Feng is with the National University of Singapore, Singapore 119077. E-mail: elefjia@nus.edu.sg.

Manuscript received 31 July 2017; revised 9 Sept. 2018; accepted 17 Sept. 2018. Date of publication 23 Sept. 2018; date of current version 11 July 2019.
(Corresponding author: Mengmi Zhang.)
Recommended for acceptance by L. Zelnik-Manor.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2018.2871688

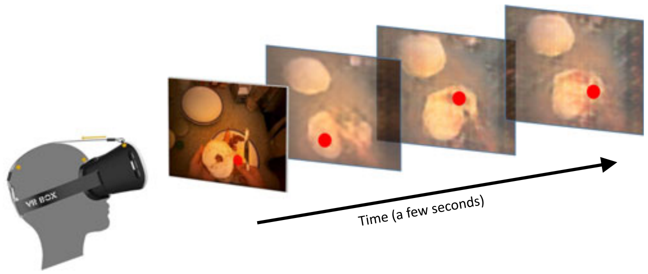


Fig. 1. Problem illustration: Gaze anticipation on future frames within a few seconds on egocentric videos. Given the current frame, the task is to predict the future gaze locations. Our proposed DFG method solves this problem through synthesizing future frames (transparent ones) and predicting corresponding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

(GAN) through a competition between a generator and a discriminator, and then predicts the gaze locations on these frames as bottom-up approach (*DFG-G*). Meanwhile, DFG anticipates the gaze prior maps as task influences (*DFG-P*) that mediates the bottom-up temporal saliency maps from the generator in *DFG-G*. Based on the latent representation extracted from the input frame before the generator, we use another 3D-CNN to predict spatial priors for gaze locations. This is the direct approach where *DFG-P* makes reasonings about the episodic steps in the task according to the semantic information extracted from the current frame without the intermediate future frame generation step. These goal-driven spatial priors bias the bottom-up saliency prediction leading to higher anticipation accuracy.

Evaluations of DFG on public egocentric datasets show that DFG boosts the performance of gaze anticipation to a considerable extent surpassing all the competitive baselines. In addition to cooking tasks, DFG demonstrates its capacity of generalizing to the object search task on Object Search Task Dataset (OST) [4]. Although DFG is not specifically trained for conventional gaze prediction problem on current frames, our GAN-based framework also significantly advances the state-of-the-arts for this problem. Moreover, we extend beyond egocentric videos and introduce the novel gaze anticipation problem on third person videos where the background is often static. In this case, DFG also achieves the best performance among all the baselines. Our rigorous analysis in the experiment section validates that our architecture can be generalized to diverse foreground and background motions. At last, we integrated our anticipated gaze locations with the existing activity recognition network. The reported results verify that anticipated gaze helps egocentric activity recognition.

In summary, our paper has the following contributions:

- We introduce a novel and important problem of gaze anticipation on egocentric and third-person videos.
- We propose an integrated framework consisting of GAN-based bottom-up stream and task-specific stream. Complementary to bottom-up approach, a task-specific mechanism estimates gaze spatial priors and biases the bottom-up saliency predication where the task information can be extracted from the current frame.
- Instead of handcrafting visual cues for gaze prediction on egocentric and third person videos, such as hands and objects, our model automatically learns these cues during end-to-end training.
- Our proposed method outperforms all the competitive baselines and demonstrates its capacity of anticipating gazes in both egocentric and third-person videos across various activities, such as cooking and object search. Without any additional training, our model also achieves state-of-the-art performance in the gaze prediction problem on current frames.

This paper extends our previously published method [4] by introducing the additional task-specific attention stream which further boosts the gaze anticipation performance. Apart from the updated experimental results on egocentric videos using the newly integrated architecture, we also introduce the novel gaze anticipation problem on third person videos and provide evaluation results on public datasets. Moreover, we add more experimental investigations about our architecture design by exploring the potential factors influencing gaze anticipation performance and comparing the ablation results on both egocentric and third person videos.

2 Preliminaries and Related Work

We review important works related to computational models of visual attention and gaze prediction on egocentric and third-person videos. As our method is inspired by generative video models, we also provide literature reviews on video generation approaches in computer vision.

2.1 Saliency Prediction

Computational saliency models are based on feature-integration theory [11] where low-level features, such as color, contrast and intensity, are combined. The first models were developed by Koch et al. [12] and Itti et al. [13]. Subsequent works [14], [15], [16] further improve saliency map predictions via various methods such as graph-based saliency model [14], boolean map based saliency [17] and information maximization-based saliency [18]. With the increasing availability of larger scale human fixation datasets, a number of works [19], [20], [21] employed the data-driven approaches. These works explored the best feature combinations from a set of low and high level features, such as objects and scene context, using support vector machine (SVM) [22], [23], least-square regression and AdaBoost [24].

The most recent saliency models leverage rich pools of semantic regions or objects in the scene from deep convolutional neural network [25], [26]. The first attempt to leverage deep learning for saliency prediction was [27] where they used the response from convolution layers as feature maps to classify the fixated regions. Subsequent work [28] was developed based on Alexnet [29] initially for object recognition network. In [30] and [25], deep models were applied across coarse and fine scales and then a large number of other neural network models emerged in saliency community; but they focus on image saliency prediction and the motion information across frames has been discarded.

2.2 Gaze Prediction on Videos

In egocentric video analysis, Ba et al. [31] proposed to analyze visual attention by exploring correlations between head orientation and gaze direction. Similarly, Yamada et al. [2] presented gaze prediction models and explored motion correlations with the aid of external motion sensors. Borji et al. [32] 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. The most recent model on gaze prediction in hand-object manipulation tasks was proposed by Yin et al. [3]. Hand detection and pose recognition provide primary egocentric cues in their model. Since their egocentric cues are predefined, their model may not generalize well to various egocentric activities especially when hands are not involved.

There is rich literature in gaze prediction on third-person videos, such as [33], [34], [35], [36]. Most of these works model temporal dynamics by salient candidate selection across time [35], space-time whitening [34] or video compression [33]. One of the recent works [36] is developed based on Long Short-Term Memory (LSTM) which learns the essential spatial-temporal features via end-to-end training. However, it is not clear how these methods could perform on egocentric videos. Moreover, the inputs to these frameworks often require multiple frames, whereas the input in our gaze anticipation problem is one single current frame.

2.3 Generative Video Models

Learning how scenes transform with time is a fundamental research problem. Our work builds upon state-of-the-art adversarial learning methods [37], [38]. Generative adversarial networks have shown fascinating performance for image modeling [37], [39], [40] and we extend to videos. Notably, [41] and [42] also use generative adversarial networks for video frame prediction. [42] improves the video generation performance in terms of predicting videos for longer time scale and learning video prediction using unlabeled data. There are also related works in video generation [43], [44], [45], [46] and future perdition [47], [48], [49], [50]; however, most of these generative models are conditioned on the past frames. The recent work [51] proposes to use biologically-inspired predictive coding mechanism to learn temporal scene dynamics by providing both feed-forward and feedback connections. Different from their work, we use spatial-temporal networks to jointly generate a sequence of future frames which prevents error accumulating due to iterative feeding the generated frames back as the input. Our work is also related to a large body of works that applies spatial temporal networks on unlabeled videos for visual recognition tasks [52], [53], [54], [55]. Instead, we adapt them for video generation and hence, gaze anticipation. To the best of our knowledge, we are the first to tackle gaze anticipation problem on both egocentric videos and third-person videos.

3 Our Model

In this section, we first introduce an overview of our proposed model, Deep Future Gaze, 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 an integrated framework consisting of two pathways: task-specific pathway *DFG-P* and bottom-up pathway *DFG-G* as shown in Fig. 2. In *DFG-G*, it consists of two modules: generative adversarial networks-based *Future Frame Generation* and *Temporal Saliency Prediction*. In *Future Frame Generation*, it has two networks: *Generator* and *Discriminator*.

Generator generates future frames and then *Temporal Saliency Prediction* predicts their corresponding temporal saliency maps, i.e., spatial probabilistic maps of gaze locations across time. *DFG-G* is regarded as the bottom-up pathway where the attention is driven by external stimuli (the generated future frames). Complementary to *DFG-G*, we add in *DFG-P* to estimate the priors of gaze locations without the intermediate future frame generation step. It makes inference about the gaze distribution in the task at hand based on the latent representation of the input frame. In the end, the task-specific attention mechanism from *DFG-P* mediates the bottom-up attention in *DFG-G*. The temporal saliency maps predicted from *DFG-G* get biased by the gaze spatial priors via element-wise summation. The spatial coordinates with the maximum probability are output as the anticipated gaze locations.

3.2 The Generator Network

In *Future Frame Generation*, the goal of *Generator* 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. Hence, $I_{t+1,t+N}$ can be used for predicting N temporal saliency maps $S_{t+1,t+N}$ in *Temporal Saliency Prediction*. 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, *Generator* 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 4.12.

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

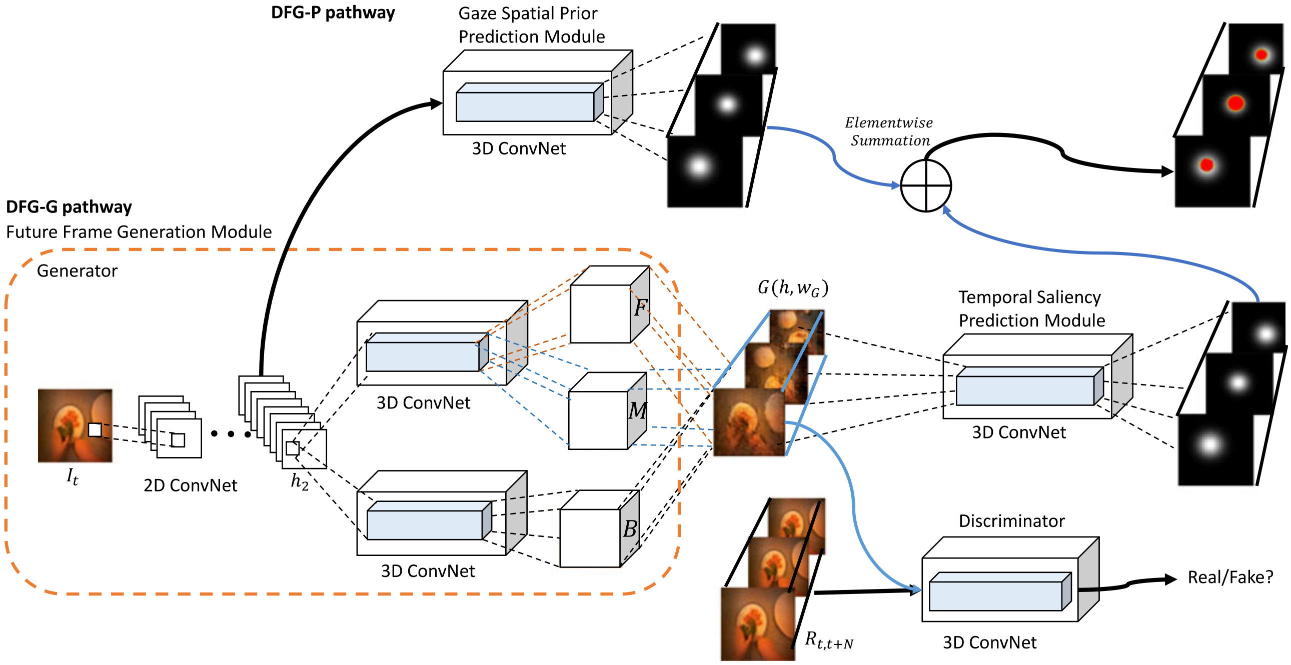


Fig. 2. Architecture of our proposed deep future gaze (DFG) model. It contains *DFG-P* and *DFG-G* pathways. In *Generator* in *Future Frame Generation Module* in *DFG-G*, latent representation of the current frame I_t is extracted by 2D ConvNet. To explicitly untangle foreground and background, it then 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 temporal saliency maps. As a competitor to *Generator*, *Discriminator* 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. *DFG-P* predicts the gaze spatial priors from the task at hand inferred from the latent representation of I_t. Element-wise summation is performed on the gaze spatial prior maps and the temporal saliency maps to produce the anticipated gaze locations (red dots).

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.

We note that, in egocentric videos, there often exists a clear distinction between foreground and background motions. While foreground objects tend to move together more coherently among themselves, they tend to distinguish from background objects due to motion relativity. For example, when the subject is transferring the food in hands from one place to another, foreground objects, such as arms and manipulated objects, tend to be always in the center of the egocentric frames while the background objects are moving in the opposite direction of head movements in the egocentric frames. The coherence within foreground and background motions themselves and the clear boundary between these motions make DFG learn to distangle the foreground objects from the background automatically during frame generation even though there is no specific training loss to explicitly supervise the network to distinguish these two.

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, *Generator* is followed by *Temporal Saliency Prediction* to generate temporal saliency maps of dimension $N \times 1 \times W \times H$.

3.3 The Discriminator Network

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 *Discriminator* as a competitor to *Generator*, by providing the additional feedbacks to *Generator* [56].

Discriminator 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, *Discriminator* follows the same architecture as the foreground generation model other than replacing all the upsampling layers with the convolution layers and this architecture has also been shown to be effective in [56]. The output is a binary label indicating whether the input frame is fake or real.

3.4 DFG Gaze Spatial Prior Pathway (DFG-P)

As a complementary of *DFG-G* pathway, *DFG-P* estimates the gaze spatial priors based on the latent representation $h(I_t)$ of the current frame I_t in *Generator*. The semantic information in $h(I_t)$ underlying the task information contributes to the inference about the distribution of gaze locations in the next few seconds. To ensure the gaze movements to be coherent across spatial and temporal domains, we use a 3D-CNN in *DFG-P* to estimate the prior maps for gaze locations of dimension $N \times 1 \times W \times H$. At the training stage, the 3D-CNN encodes the spatial distributions of gaze locations and their motion trajectories corresponding to the episodic steps in the task at hand.

In the end, the gaze prior maps from *DFG-P* mediate the temporal saliency maps from *Temporal Saliency Prediction* module. The bias from the task information is fused with the stimuli-driven bottom-up attention mechanism via an element-wise summation operation. We normalize the spatial prior maps and the temporal saliency maps to be within range $[0, 1]$ before element-wise summation. Concerned with the large variance of gradient changes in element-wise multiplication, we use element-wise summation instead to adaptively tune the effect of the task-specific bias on the bottom-up saliency. The results after element-wise summation are normalized again and the highest activation points on these probabilistic maps are the most probable anticipated gaze locations.

We should be cautious that, there is no top-down modulation in DFG. *DFG-P*, which carries task-specific information, is still a feed-forward 3D-CNN. Complementary to realistic visual features that guides gaze anticipation in *DFG-G*, *DFG-P* relaxes constraints on visual features and learns task-specific gaze priors or any abstract representations of the task useful for gaze anticipation. For example, in "spreading jam on bread" task, given the current frame showing the human subject puts the bread on the plate which is probably in the lower half of the egocentric view, *DFG-P* predicts high attention values to the upper half of the egocentric view (the table where all bottles are located) in the next few seconds due to the "jam bottle grabbing" task while *DFG-G* estimates the visual saliency of all bottles on the table and selects the jam-bottle like visual features.

3.5 Training

Training. We train DFG end-to-end by stochastic gradient descent with learning rate 0.00005 and momentum 0.5. Adam Optimizer [57] is used. *Generator* and *Discriminator* play against each other. *Generator* is designed to predict future frames as "real" as possible to fool *Discriminator*, while *Discriminator* 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 *Discriminator* 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 *Discriminator*.

As the opponent of *Discriminator*, *Generator* needs to satisfy two requirements: 1) the generated outputs should be real enough to fool *Discriminator*; 2) the initial output of the generated frames should be visually consistent with the current frame. The objective function for training *Generator* 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 [41].

Temporal Saliency Prediction takes $I_{t+1,t+N}$ as input to generate temporal saliency maps. *Temporal Saliency Prediction* 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. The fixation map refers to the binary map where we use 1 to indicate the human gaze location. To avoid sparseness of fixation maps, we convolve each binary fixation map with a gaussian mask and then we normalize it to be within range $[0, 1]$.

Similarly, *DFG-P* takes the latent representation $h(I_t)$ of the current frame I_t as the input to generate gaze spatial prior maps. We train *DFG-P* in a supervised manner using the same KLD loss function in Equation (5) where P_i is the temporal fixation map and Q_i is the gaze spatial prior map for the $(t + i)$th frame.

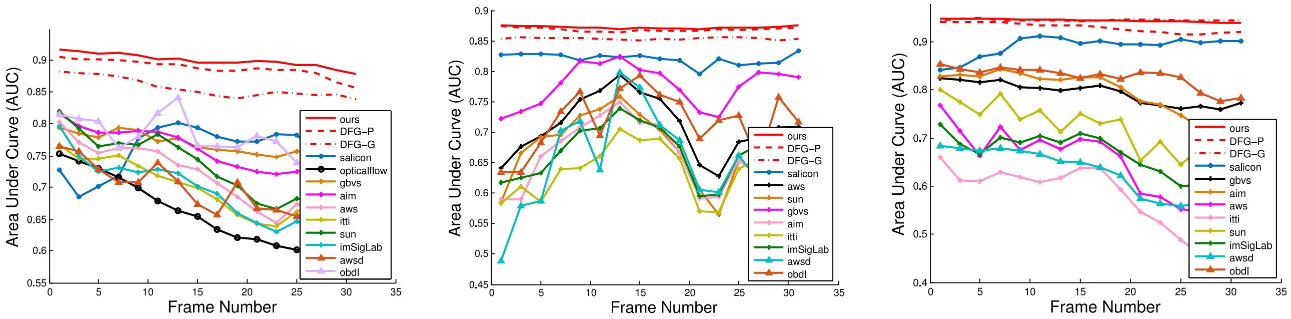
3.6 Implementation Details

DFG is developed based on [56] in Torch. The source code is available at https://github.com/Mengmi/deepfuturegaze_gan. We train everything from scratch 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 $[-1, 1]$. The gaze spatial prior maps and the temporal saliency maps are of the same dimensions where $N = 32$, $W = 64$, and $H = 64$.

*Gaze prediction on current frame*DFG can also be used for gaze prediction on the current frame. Since *Generator* 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 all public datasets using standard evaluation metrics. We also provide detailed analysis of



(a) GTEA Dataset (2.1 sec ahead) (b) OST Dataset (3.2 sec ahead) (c) Hollywood2 Dataset (1.3 sec ahead)

Fig. 3. Evaluation of gaze anticipation using area under the curve (AUC) on the current frame as well as 31 future frames in GTEA, OST and Hollywood2 dataset. Evaluation results in GTEAplus dataset are similar as GTEA. See Supplementary Material, available online, for evaluation results of gaze anticipation in GTEAplus Dataset. Larger is better. The algorithms in the legend are introduced in Section 4.3.

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 [58]. 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 subjects are asked to prepare meals freely. Same as Yin et al. [3], we use videos 1, 4, 6-22 as training set and the rest as test set.

GTEAplus Dataset [3]. 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 [3].

Object Search Tasks. To explore whether DFG can be generalized well for other tasks in egocentric contexts, we include the public egocentric video dataset in object search [4]. This dataset consists of 57 sequences on search and retrieval tasks performed by 55 subjects in a fully furnished and functional model home. 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 and move them to the packing location (dining table). Compared with GTEA and GTEAplus, this dataset involves larger head motions and the human subjects have to walk around and look for objects in the search list with hands appearing less frequently.

Hollywood2 Dataset [59]. This is a public third person video dataset with 12 classes of human actions. Mathe and Sminchisescu [60] provides the gaze data for this dataset to study gaze dynamics. We include a subset of this dataset to evaluate DFG on gaze anticipation in the context of third person videos. In particular, video clips with these four actions related to social interactions are included in our experiment: handshaking, person hugging, kissing and person fighting. Among 3,669 video clips in total, there are 365 video clips for training and 127 for testing and validation.

4.2 Evaluation Metrics

We use four standard evaluation metrics on gaze anticipation: Area Under the Curve (AUC) [61], Average Angular Error (AAE) [62], Normalized Scanpath Saliency (NSS) [63] and Precision-Recall Curve (PR) [64] as below.

Area Under the Curve 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 is the angular distance between the predicted gaze location and the ground truth.

Normalized Scanpath Saliency computes the average normalized saliency at the fixated locations.

Precision-Recall Curve represents results for binary decision in machine learning [64]. We report the area under the precision-recall curve at the ith future frame.

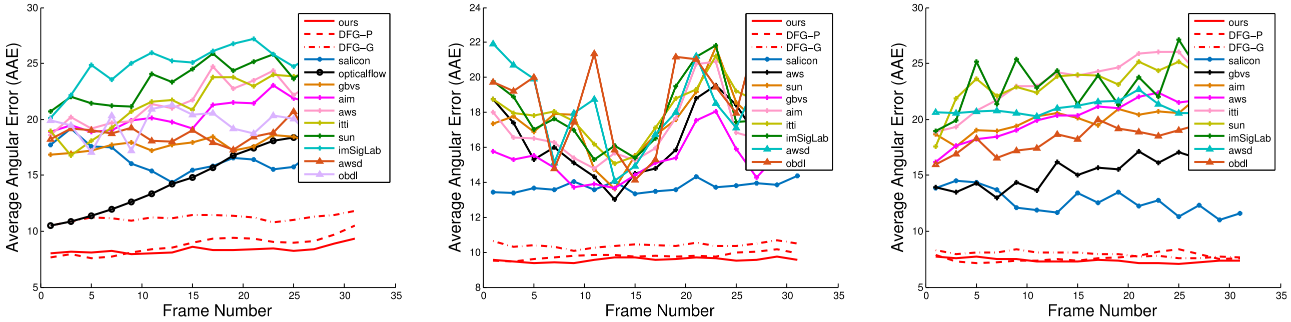
There are four datasets with four evaluation metrics resulting in 16 combinations. We report the gaze anticipation evaluation results in full using *all* evaluation metrics across *all* four datasets. For simplicity, in ablation study and architecture analysis, we opt to focus on reporting the analysis results on GTEA in egocentric videos and Hollywood2 in third person videos as representatives only using AUC and AAE in the main text. Refer to the Supplementary Material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2018.2871688, for some evaluation results on other datasets. For consistency, except for Figs. 3 and 4 where we show the metrics scores for all future 31 frames, we report the *mean* gaze anticipation accuracy by *averaging* the metrics scores over the current frame as well as the next 31 future frames.

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 *Generator* 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) [14], Natural Statistics Saliency (SUN) [15], Adaptive Whitening Saliency (AWS) [65], Attention-based Information Maximization (AIM) [66], Itti's Model (Itti) [67], and Image Signature Saliency (ImSig) [68]. Moreover, we also include gaze prediction methods on videos [34] (AWSD) and [33] (OBDL).

Second, SALICON [25] 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 pretrained SALICON model is tested on our generated frames for gaze anticipation.



(a) GTEA Dataset (2.1 sec ahead) (b) OST Dataset (3.2 sec ahead) (c) Hollywood2 Dataset (1.3 sec ahead)

Fig. 4. Evaluation of gaze anticipation using Average Angular Error (AAE) on the current frame as well as 31 future frames in GTEA, OST and Hollywood2 Dataset. Evaluation results in GTEAplus dataset are similar as GTEA. See Supplementary Material, available online, for evaluation results of gaze anticipation in GTEAplus Dataset. Smaller is better. The algorithms in the legend are introduced in Section 4.3.

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 [69]. 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 [3] for gaze prediction on current frames in GTEA and GTEAplus. We exclude this method on OST since the required hand annotations by [3] are not available. We also cannot extend this method to gaze anticipation problem.

4.4 Results of Gaze Anticipation on Egocentric Videos

DFG surpasses all the competitive baselines significantly in gaze anticipation in egocentric videos. We report the quantitative evaluation results in Fig. 3 (AUC), Fig. 4 (AAE) and Table 1 (NSS and PR) on egocentric datasets.

Over all egocentric datasets (GTEA, GTEAplus, and OST), DFG outperforms all the competitive baselines. In particular, we observe a significant performance boost with respect to our previous method (*DFG-G*) [4] which is the second best as shown in Figs. 3 and 4 by 26.2, 12.0 and 8.8 percent in relative advance (RA) in AAE and 4.5, 0.05 and 2.3 percent in RA in AUC. RA in percentage is computed as

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

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 *DFG-G* on the ith future frame. Complementary to *DFG-G*, fusion with *DFG-P* greatly improves the gaze anticipation performance which emphasizes the necessary role of *DFG-P* pathway which predicts gaze priors for the task at hand and biases the saliency maps predicted by *DFG-G*. See Section 4.9 for more analysis.

TABLE 1
Averaged Gaze Anticipation Performance over Current Frame as well as 31 Future Frames Using Normalized Saliency Scanpath (NSS) and the Area Under the Precision-Recall Curve (PR)

	GTEA		GTEAplus		OST		Hollywood2	
Metrics	NSS	PR	NSS	PR	NSS	PR	NSS	PR
ours	**1.62**	**0.50**	**1.95**	**0.53**	1.45	**0.48**	**1.91**	**0.56**
SAL [25]	0.97	0.46	1.11	0.43	**1.91**	0.45	1.76	0.49
GBVS [14]	0.94	0.42	1.52	0.44	0.75	0.43	0.54	0.41
AWS [65]	0.73	0.39	0.74	0.42	0.13	0.39	−0.05	0.41
AIM [66]	0.91	0.39	0.85	0.39	0.55	0.42	0.73	0.41
SUN [15]	0.77	0.38	1.58	0.46	0.74	0.41	0.65	0.38
Itti [67]	0.67	0.40	1.01	0.40	0.18	0.43	−0.22	0.41
ImSig [68]	0.62	0.38	1.03	0.39	0.40	0.42	0.56	0.41
AWSD [34]	0.69	0.40	1.06	0.42	0.56	0.41	0.44	0.41
OBDL [33]	1.02	0.42	1.21	0.42	0.78	0.42	1.14	0.44

Higher is better for NSS and PR. Best results are in bold.

Qualitative results in Fig. 5 demonstrate that DFG learns to untangle foreground and background motions. For example, both the hand and the object (the bun) get highlighted in the foreground. 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.

It is also observed that the temporal saliency maps anticipated by *DFG-P* and *DFG-G* are visually different. Though *DFG-P* assigns high attention values to the manipulation point (slightly below the center of the egocentric field of view across all future frames in general during the table-top food preparation process), it fails to capture the hand motion when the subject is rotating the bun within the local region; conversely, *DFG-G* anticipates the effect of local hand motion and hence, predicts slight attention shifts in the future frames. More qualitative results in Supplementary Material, available online, demonstrate that *DFG-G* and *DFG-P* can be jointly adapted in different tasks which cover varieties of illumination conditions, head orientations, hand poses, and manipulated objects.

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 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 provide comparisons with gaze prediction methods on videos [33], [34]. Although these methods take temporal

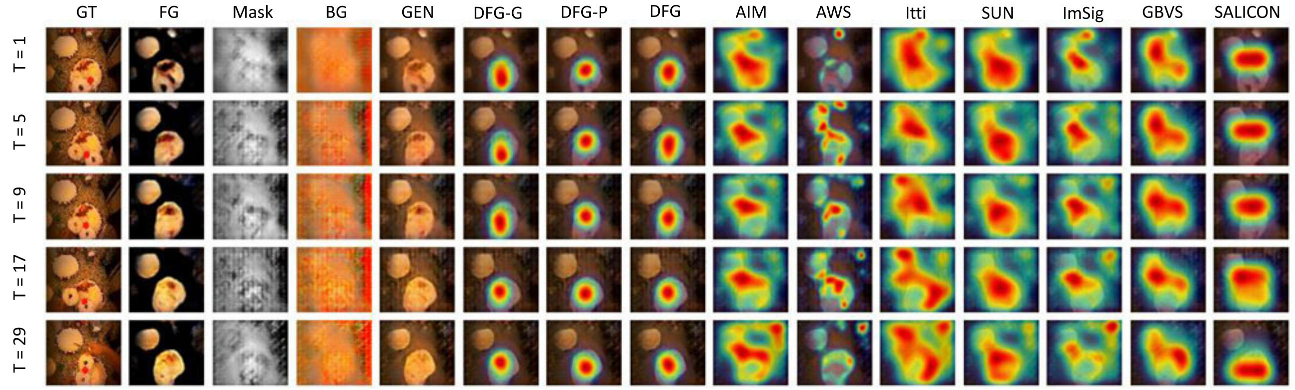


Fig. 5. Example results of gaze anticipation on GTEAplus egocentric video dataset. Our DFG model produces 31 future frames based on the current frame. From first to last rows, results on future frames #1, 5, 9, 17, 29 with respect to the current frame are shown. The leftmost column shows the ground truth (GT) with red circle denoting human gaze locations. Column 2, 3, 4 (FG, mask, BG) show the foreground $F(\cdot)$, the mask $M(\cdot)$, and the background $B(\cdot)$ learnt by *Generator* respectively. Column 5 shows the generated future frames (GEN). Column 6 and 7 show the corresponding predicted temporal saliency maps from two pathways *DFG-G* and *DFG-P* in our model. Column 8 show the final integrated temporal saliency maps predicted by our model. Column 9 and onwards show the predicted temporal saliency maps by all baselines (See Section 4.3). Best viewed in color. See Supplementary Material, available online, for more qualitative examples.

information into account, these feature cues (space-time whitening and information from video compressors) on synthetic frames are still not sufficient compared with *DFG-G* [4]. Another missing element in these models is task-specific information which is also critical for gaze anticipation.

4.5 Results of Gaze Anticipation on Normal Videos

Beyond egocentric videos, we test DFG on third person videos where the backgrounds are often static. From the quantitative evaluation results in Fig. 3c (AUC), Fig. 4c (AAE) and Table 1 (NSS and PR), DFG achieves the best performance in Hollywood2 dataset with four evaluation metrics. Using Equation (6), DFG outperforms our previous method (*DFG-G*) [4] by 7.1 percent in relative advance in AAE and 0.09 percent in RA in AUC.

We present a qualitative example in Fig. 6 in hand shaking scenario in Hollywood2. From the results, it demonstrates that DFG is also capable of segmenting foreground objects from static backgrounds in third person videos. For example, the three persons get highlighted in the mask. As the background is uniform over time, this is reflected in the darker regions of the mask as well as the bright regions in the background stream. Furthermore, we also observe that DFG can adaptively generate “realistic” future frames regardless of variant color conditions, such as the gray-scale video frames as shown in Fig. 6.

We also note that *DFG-P* learns the general gaze anticipation patterns when it requires complex gaze shifts while human subjects are observing a video clip in a social interaction task. The qualitative example in Fig. 6 shows an occasion where three persons are having a conversation. Though there is no significant visual change in this social interaction case and *DFG-G* predicts almost static future frames over time, *DFG-P* anticipates attention spread across the three persons where the highest activation points on the saliency maps shift from the center to the left across frames which is consistent with the ground truth gaze patterns.

Compared with the performance on egocentric videos, SALICON performs relatively better on third person videos. This is because the backgrounds in video clips in Hollywood2 are often static which alleviates the demands of temporal information. In addition, the semantic information such as faces appear often in social interaction tasks where SALICON is good at attending to these semantic objects on each frame. The performance of the rest of the baselines on Hollywood2 is consistent with those in egocentric videos.

4.6 Spatial Bias Analysis

In this section, we study the various spatial biases including center bias, gaze fixation distribution from the training data as well as head motion and how they may effect the gaze anticipation performance in egocentric and normal videos.

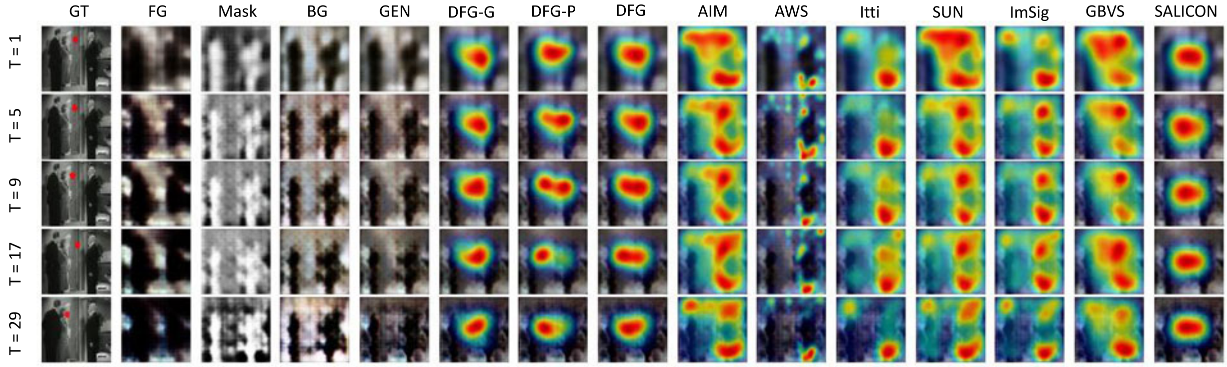


Fig. 6. Example results of gaze anticipation on Hollywood2 third person video dataset. The format and conventions follow those in Fig. 5.

TABLE 2
Evaluation of Center Bias Effect over the Next 31 Frames

sAUC	GTEA	GTEAplus	OST	Hollywood2
DFG(ours)	**0.62**	**0.57**	**0.57**	**0.52**
Center Bias	0.5	0.5	0.49	0.49

TABLE 3
Average Spatial Bias and Human Performance over the Next 31 Frames on GTEA and GTEAplus Datasets

	GTEA		GTEAplus	
	AUC	AAE	AUC	AAE
Our Best	**0.90**	**8.3**	**0.94**	**5.9**
GazeDistriMap	0.86	9.3	0.93	7.4
GazeDistriMap + DFG-G	0.88	9.0	0.94	6.8
Human	0.66	9.5	0.77	6.8

4.6.1 Center Bias

We often observe a strong center bias in egocentric videos. This is due to the fact that egocentric videos are captured from the first person view. 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. Similarly, center bias is also present in free-viewing tasks in static images and third person videos [70]. As AUC favors center bias, we use shuffled-AUC (sAUC) to compare our model with center bias and we report its sAUC score in Table 2. It confirms that our model learns to anticipate gaze by taking various semantic information and motion dynamics into account instead of predicting center bias on future frames over all datasets.

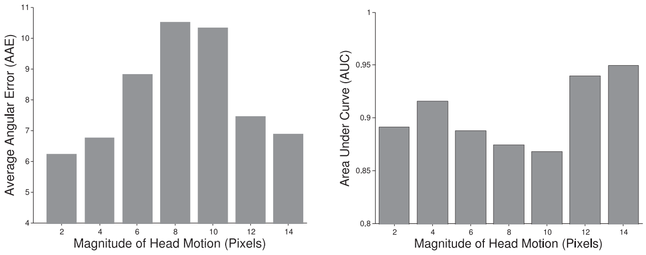
4.6.2 Gaze Distribution Map

We report the two variations of utilizing the 2D gaze distribution map computed from all human fixations in the training set: (1). the 2D gaze distribution map alone as the predicted temporal saliency map on all future frames; (2) we replace *DFG-P* in our DFG model with the gaze distribution map. See Supplementary Material, available online, for implementations of these two variants.

Table 3 shows the gaze distribution map alone (Row 2) is much worse than our DFG model (Row 1). Though *DFG-G* with gaze distribution map (Row 3) is better than gaze distribution alone, it is still inferior to DFG by 1 in GTEA and 1.5 in GTEAplus in terms of AAE. This suggests the gaze prior has complex dynamics and *DFG-P* which learns gaze prior variations depending on the task specifications is important for gaze anticipation.

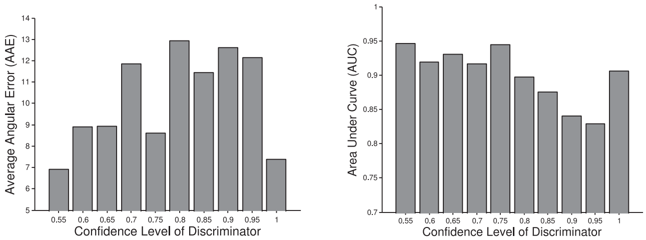
TABLE 4
Statistics of Camera and Gaze Motions

	Gaze Motion			Camera Motion		
	Mean	Median	Variance	Mean	Median	Variance
GTEA	20.4	13.5	508	6.7	3.6	92
GTEAplus	7.1	5.0	89	9.9	5.8	135



(a) Average Angular Error (b) Area Under the Curve

Fig. 7. Evaluation of average gaze anticipation performance over 31 future frames versus magnitude of head motions in GTEA.



(a) Average Angular Error (b) Area Under the Curve

Fig. 8. Evaluation of average gaze anticipation performance over 31 future frames versus confidence of *Discriminator* in our model in GTEA.

4.6.3 Head Motion

We provide the statistics of head and gaze motion in pixels in our test data in GTEA and GTEAplus datasets. As there is no ground truth for head motion, we estimate it by averaging the dense optical flow in the boundary pixels between adjacent frames. With respect to a frame (480 by 640 in pixels), the statistics of amplitudes for these motion are reported in Table 4. To study the effect of head motion on gaze anticipation, we calculate the averaged magnitude of head motion across the next 31 ground truth frames and report the averaged gaze anticipation performance on these frames in Fig. 7. In general, the gaze anticipation performance of our DFG model drops when there is larger head motion (see Supplementary Material, available online, for qualitative examples). However, its performance does not monotonically decrease. It is possible that when there is a very large head motion, gaze shift gets compensated and aligns with head orientations. Due to the complex nature and large variances between gaze and head motions, our analysis confirms that the two-stream *Generator* in our DFG model is critical for better gaze anticipation by estimating the these two motions separately.

4.7 Discrepancy of Future Frames from Real Scenes

We study how discrepancy of the future frames from the real scene will effect gaze anticipation performance. To quantitatively evaluate the quality of the generated future frames from *Generator*, we compute the confidence of *Discriminator* which acts as a competitor against *Generator* striving to distinguish whether the generated frames are real or synthetic. The more confident *Discriminator* is, the easier for *Discriminator* to tell real ones from the synthetic; hence, the more discrepancy there is between the generated future frames generated by *Generator* and the real scene. Ideally, if the synthetic frames are indistinguishable from real frames, the *Discriminator* confidence is 0.5. Fig. 8 shows the average gaze anticipation performance over the next 31 future

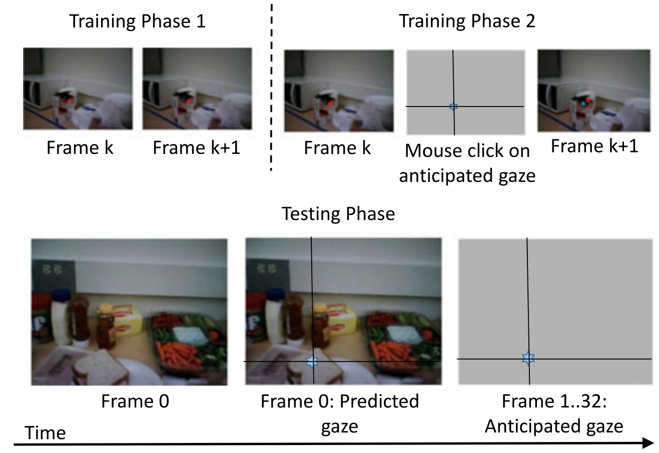


Fig. 9. Schematic description of human psychophysics experiment on gaze anticipation. In Training Phase 1, subjects are presented with all the video frames of 5 training video clips and their corresponding overlaid ground truth gaze locations denoted by red circles. In Training Phase 2, subjects are first presented with the current video frame with ground truth gaze location same as Training Phase 1 followed by a blank gray screen. Subjects use computer mouses to click on the anticipated gaze location for the $t+1$th frame. Next, the ground truth video frame overlaid with ground truth gaze location (red circle) and mouse click location (blue cross) are shown. Repeat for all 5 training video clips. In the testing phase, subjects are only presented with the current frame. They have to use computer mouse to click on the predicted gaze location on the current frame as well as anticipated future gaze locations on blank gray screen for a total of 100 testing video clips (50 clips per dataset in GTEA and GTEAplus).

frames versus the confidence of *Discriminator*. The gaze anticipation performance is positively correlated with the quality of the generated frames which validates that *Discriminator* is critical for providing feedbacks to *Generator* in order to generate more realistic future frames useful for improving gaze anticipation performance.

4.8 Human Performance on Gaze Anticipation

As human benchmark is a gold standard in many computer vision tasks and it is not clear how humans perform in our gaze anticipation task, we conduct human psychophysics experiments to test human performance in this task. For fair comparison with the computational models, we provide 4 human subjects (22-28 years old, 2 females, 2 males) with two training phases and test them on gaze anticipation tasks on 50 video clips per test set from GTEA and GTEAplus datasets. See Fig. 9 for experiment schematics and Supplementary Material, available online, for detailed description of experimental procedures.

We report the average human performance on gaze anticipation task over the next 31 future frames in Table 3, Row 4. Human performance is as good as gaze fixation maps with *DFG-G* but still inferior to our DFG model. However, this result cannot be over-interpreted as there are several differences between humans and the computational models: (1) number of training samples (humans are exposed to fewer training samples compared with DFG); and (2) knowledge of the tasks (humans do not have full knowledge about all the task information in each dataset while computational models are trained with more varieties of tasks). This is an interesting future research direction and it suggests promising real life applications where the computational models could assist humans in several domains involving gaze anticipation, such as health care and autonomous driving.

TABLE 5
Ablation Study on GTEA, OST and Hollywood2 Datasets

	GTEA		OST		Hollywood2	
	AUC	AAE	AUC	AAE	AUC	AAE
Our Best (DFG)	**0.90**	**8.3**	**0.87**	**9.5**	**0.95**	**7.4**
DFG-P	0.88	8.9	0.87	9.8	0.93	7.5
DFG-G	0.86	11.3	0.85	10.3	0.94	7.9
One-stream	0.85	12.0	0.86	10.5	0.95	7.7
Replace(GT)	0.82	13.5	0.80	13.0	0.86	12.6
Remove(D)	0.83	12.0	0.85	10.6	0.88	14.3

4.9 Ablation Study on Egocentric and Normal Videos

In order to study the effect of the individual component of DFG on both egocentric and third person videos, we do an ablation study and test on GTEA, OST and Hollywood2 datasets by removing *only* one component in DFG at one time while the rest of the architecture remains the same. There are five tests: (1) we remove *DFG-G* and evaluate the predicted temporal saliency maps from *DFG-P* only; (2) we remove *DFG-P* and this is the same as our previous algorithm with only *DFG-G* [4]. (3) we replace the two-stream 3D-CNN in *Generator* with the same structure as [56], i.e., the background stream is 2D-CNN which assumes the background is "static" while the foreground stream remains the same; (4) we train *Temporal Saliency Prediction* directly on real frames and test it on the generated frames from *Generator*; (5) we remove *Discriminator* and we only use L1 distance loss for future frame generation. Scores for gaze anticipation in AAE and AUC are averaged across future 31 frames as shown in Table 5. See Supplementary Material, available online, for schematics of the ablated models.

Compared with our previous method *DFG-G* [4], we proposed a complementary task-specific *DFG-P* and integrated it with *DFG-G*. To study its effectiveness, we test each of these two pathways individually. *DFG-P* alone performs better than *DFG-G* by 2.4 in GTEA, 0.5 in OST and 0.4 in Hollywood2 in terms of AAE but both pathways are worse than our integrated framework (DFG). We also duplicate the results of *DFG-P* (Row 2) and *DFG-G* (Row 3) in Figs. 3 and 4. We observe that both individual pathways outperform all the baselines significantly. It suggests that both the bottom-up attention mechanism *DFG-G* and the gaze prior maps predicted from task-specific information by *DFG-P* have essential contributions to gaze anticipation in egocentric and third-person videos.

4.9.1 Ablation Analysis on Egocentric Videos

The third ablation study (Row 4) on changing the background stream to a static one leads to an increase of 3.7 in GTEA and 1 in OST in terms of AAE. This implies the two-stream 3D-CNN in *Generator* is essential for learning foreground and background motions which can further improve gaze anticipation accuracy.

Compared with DFG, the fourth ablated model (Row 5) with *Temporal Saliency Prediction* trained on real frames

TABLE 6
Results of Gaze Prediction on the Current Frame

	GTEAplus		GTEA		Our OST		Hollywood	
Metrics	AUC	AAE	AUC	AAE	AUC	AAE	AUC	AAE
DFG(ours)	**0.95**	**5.6**	**0.92**	**8.1**	**0.88**	**9.6**	**0.95**	**7.75**
DFG-P	0.93	6.2	0.9	7.69	0.88	9.5	0.94	7.9
DFG-G [4]	0.95	6.6	0.88	10.5	0.85	10.6	0.95	8.3
Yin [3]	0.87	7.9	0.88	8.4	-	-	-	-
SAL [25]	0.82	15.6	0.76	16.5	0.85	13.3	0.84	14.0
GBVS [14]	0.80	14.7	0.77	15.3	0.71	18.8	0.75	10.5
AWS [65]	0.82	14.8	0.78	17.5	0.56	22.8	0.5	17.5
AIM [66]	0.76	15.0	0.82	14.2	0.77	17.0	0.75	14.4
SUN [15]	0.84	14.7	0.80	18.1	0.53	25.0	0.66	17.7
Itti [67]	0.75	19.9	0.75	18.4	0.62	19.0	0.67	26.7
ImSig [68]	0.79	16.5	0.78	19.0	0.56	24.2	0.60	20.9
AWSD [34]	0.78	16.0	0.77	18.2	0.49	21.9	0.68	20.6
OBDL [33]	0.82	19.9	0.80	15.6	0.63	19.7	0.85	16.0

performs worse with an increase of 5.2 in GTEA and 3.5 in OST in terms of AAE. In DFG, *Temporal Saliency Prediction* is attached after *Generator* for temporal saliency map prediction using end-to-end training. However, *Temporal Saliency Prediction* in the third ablated model, which are 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 *Temporal Saliency Prediction* to learn these essential features on the generated future frames.

The fifth ablation study with *Discriminator* removed (Row 6) shows an increases of 3.7 in GTEA and 1.1 in OST in terms of AAE. This demonstrates that *Discriminator* is important as the feedback to *Temporal Saliency Prediction* which provides the additional constraints such that *Generator* can generate more "realistic" future frames in longer time duration. These "realistic" future frames are critical for gaze anticipation.

4.9.2 Ablation Analysis on Normal Videos

Results in Hollywood2 dataset show DFG outperforms *DFG-G* by 0.5 and *DFG-P* by 0.1 in Hollywood2 in terms of AAE. Compared with GTEA, we observe that the task-specific influences from *DFG-P* have less impacts in Hollywood2 which is a third person video dataset. As gaze information reflects human intention and behaviors, this implies that the gazes in egocentric videos are often guided by willful plans or current goals as task-specific attentional effect. This has also been verified in the literature [71], [72].

The third ablated model (Row 4) has shown marginal effect in Hollywood2 with an increase of 0.3 in terms of AAE while there is an increase of 3.7 in GTEA dataset. As the backgrounds in Hollywood2 are often static in most cases, the 2D-CNN stream in *Generator* in the ablated model could still model the semantics on the background in normal videos. However, in GTEA, the second ablated model cannot learn complex motion dynamics in the backgrounds which leads to a significant performance drop. This further verifies the necessity of splitting *Generator* into two 3D-CNN streams in order to model the foreground and background motions in egocentric videos.

Compared with DFG, the fourth ablated model (Row 5) with *Temporal Saliency Prediction* trained on real frames performs worse with an increase of 5.2 in Hollywood2 in terms of AAE. It implies that the end-to-end training on the generated frames is equivalently important in both egocentric videos and third person videos such that *Temporal Saliency Prediction* can learn essential features on the synthesized frames.

TABLE 7
Evaluation of Gaze Anticipation on Frames at Time $t+16$ and $t+32$

Average Angular Error (AAE)				
	GTEAplus		GTEA	
Models	Ours(DFG)	SALICON	Ours(DFG)	SALICON
time $t+16$	**6.0**	11.4	**8.4**	18.4
time $t+32$	**6.5**	19.5	**9.0**	16.6

Area Under Curve (AUC)				
	GTEAplus		GTEA	
Models	Ours(DFG)	SALICON	Ours(DFG)	SALICON
time $t+16$	**0.939**	0.916	**0.891**	0.710
time $t+32$	**0.937**	0.722	**0.873**	0.767

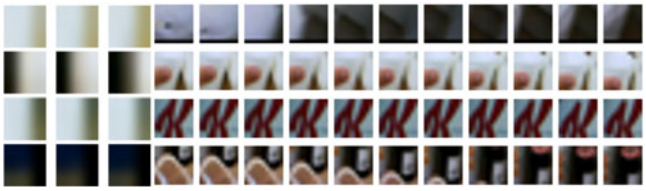
The fifth ablation study (Row 6) with *Discriminator* removed shows an increases of 6.9 in Hollywood2 in terms of AAE. This again validates the point that *Discriminator* plays a critical role in generating more realistic future frames. Moreover, we note that the performance drops more in Hollywood2 compared with GTEA. This implies that *Discriminator* is more important in the case of third person videos as the supervision from *Didscriminator* prevents over-fitting problems of *Temporal Saliency Prediction* in a more simplified task where there is less motion involved.

4.10 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 all egocentric and third person video datasets and we report both AAE and AUC scores of gaze prediction on current frames in Table 6. Number denoted in bold is the best. Results show that DFG performs better than the-state-of-the-arts even without explicitly specifying useful visual cues, such as hands, objects of interest and faces. Moreover, different from the traditional methods, our model takes the current frame as the only input without any past information. Compared with *DFG-G*, we observe that AAE scores decrease significantly and even surpass Yin et al. [3] on GTEA. It implies that the integration of task-specific information from *DFG-P* with *DFG-G* contributes to gaze prediction on current frames.

4.11 Analysis on Temporal Dependency of Gaze States

It is 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 Table 7 for results in terms of AUC and AAE on GTEA and GTEAplus. Number denoted in



(a) GP1 (b) GP4

Fig. 10. 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. (a) The filters in the first 3D convolution layer show low-level features, such as edges. (b) The regions of salient objects are highly activated in the second last convolution layer, such as the fonts on the oatmeal box.

bold is the best. 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.

4.12 Analysis on Frame Numbers

In video analysis, the number of consecutive frames is a key parameter in practice. To study the effect of the number of frames on which we anticipate gaze, we assign the scalar weights to tune the losses in both *Generator* and *Temporal Saliency Prediction* for the next 32 frames while maintaining the same architecture. See Supplementary Material, available online, for implementation details and reported results. From the results, we observe that given an input frame, in order to anticipate gazes on subsequent L frames, models trained with $L + K$ frames will perform better as K increases. This is because *Temporal Saliency Prediction* can learn the temporal dynamics with more information flowing back from the future K frames.

4.13 Visualization of Convolution Filters

As *Temporal Saliency Prediction* estimates temporal saliency maps based on the generated frames, we analyze the learnt convolution filters in *Temporal Saliency Prediction* and align the observations with human bottom-up visual attention mechanism. See Supplementary Material, available online, for visualization method of convolution filters. We observe that the filters in the first convolution layer of *Temporal Saliency Prediction* learn the low level features, such as edges and regions of high contrast. This observation aligns well bottom-up visual attention which is driven by low level features at the initial stage according to [11]. 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 Fig. 10a) and the brightness in the bottom regions decay across time (row 4 in Fig. 10a). 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 Fig. 10b. 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 butter. Overall, we infer that *DFG-G* not only learns egocentric cues in the spatial domain but also motion dynamics in the temporal domain.

4.14 Gaze-Aided Egocentric Activity Recognition

Recent papers have shown that visual attention could help in egocentric activity recognition [3], [73]. To verify our proposed future gaze model is also useful for egocentric activity recognition, we integrate gaze information into the feedforward 3D-CNN for egocentric activity recognition. As [74] shows that 3D-CNN can be used for activity recognition, we adapt the down-scaled framework from [74] (C3D) and integrate the anticipated gaze into the network. See Supplementary Material, available online, for implementation details and activity recognition accuracies. From the results, one can observe that our gaze-aided model surpasses C3D network [74] and several traditional methods [75], [76] and the guess-at-random basline significantly. By comparing the model with our predicted gaze and the one with the center gaze, it can be found that more accurate gaze prediction could result in better egocentric activity recognition. However, the wrong gaze information may be misleading for the network, which may result in poor performances as the baseline uses the center bias.

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 both egocentric and third person videos. We develop an integrated framework, named as Deep Future Gaze, consisting of two pathways: bottom-up pathway *DFG-G* built upon Generative Adversarial Network and task-specific pathway *DFG-P* generating gaze spatial prior maps which modulate the bottom-up saliency prediction. We evaluate our integrated model using standard metrics and our performance surpasses all the competitive baselines significantly in both egocentric and third-person videos covering various activities, such as cooking and object search tasks. Moreover, we investigate the potential factors contributing to better gaze anticipation performance and justify the importance of the individual component in our proposed architecture. Though our model is not specifically trained for gaze prediction problem on current frames, DFG performs better compared with the state-of-the-art. Different from all the existing methods, DFG does not require explicit egocentric cues or any past information.

Acknowledgments

This work was supported by the Reverse Engineering Visual Intelligence for cognitiVe Enhancement (REVIVE) programme (1335H00098) funded by 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, Sayed Hossein Khatoonabadi, and Victor Leboran for their help in replicating the experimental setups in [3], [33], [34].

References

[1] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, “The evolution of first person vision methods: A survey,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 744–760, May 2015.

[2] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Attention prediction in egocentric video using motion and visual saliency," in *Proc. Pacific-Rim Symp. Image Video Technol.*, 2011, pp. 277–288.

[3] Y. Li, A. Fathi, and J. M. Rehg, "Learning to predict gaze in egocentric video," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 3216–3223.

[4] M. Zhang, K. T. Ma, J. H. Lim, Q. Zhao, and J. Feng, "Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3539–3548.

[5] 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 Int.*, San Diego, CA, vol. 1, p. 8055, 2009.

[6] M. Kumar, T. Winograd, A. Paepcke, and J. Klingner, "Gaze-enhanced user interface design," Stanford InfoLab, 2007.

[7] R. Ohme, M. Matukin, and B. Pacula-Lesniak, "Biometric measures for interactive advertising research," *J. Interactive Advertising*, vol. 11, no. 2, pp. 60–72, 2011.

[8] F. Multon, L. France, M.-P. Cani-Gascuel, and G. Debunne, "Computer animation of human walking: A survey," *J. Vis. Comput. Animation*, vol. 10, no. 1, pp. 39–54, 1999.

[9] R. C. Zeleznik, A. S. Forsberg, and J. P. Schulze, "Look-that-there: Exploiting gaze in virtual reality interactions," Brown Univ., Providence, RI, USA, Tech. Rep. CS-05, 2005.

[10] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration," *Frontiers Psychology*, vol. 6, 2015, Art. no. 1049.

[11] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[12] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*. Berlin, Germany: Springer, 1987, pp. 115–141.

[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[14] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[15] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 32–32, 2008.

[16] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.

[17] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 153–160.

[18] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, pp. 5–5, 2009.

[19] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 2106–2113.

[20] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 438–445.

[21] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 28–28, 2014.

[22] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, no. Aug., pp. 1871–1874, 2008.

[24] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 9–9, 2011.

[25] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 262–270.

[26] Y. Lin, S. Kong, D. Wang, and Y. Zhuang, "Saliency detection within a deep convolutional architecture," in *Proc. Workshops 28th AAAI Conf. Artif. Intell.*, 2014.

[27] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2798–2805.

[28] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet," arXiv:1411.1045, 2014.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[30] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 362–370.

[31] S. O. Ba and J.-M. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 101–116, Jan. 2011.

[32] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 470–477.

[33] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5501–5510.

[34] V. Leboran, A. Garcia-Diaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 893–907, May 2017.

[35] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1147–1154.

[36] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," arXiv:1603.08199, 2016.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[38] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv:1511.06434, 2015.

[39] E. L. Denton, S. Chintala, R. Fergus, et al., "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[40] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 318–335.

[41] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," arXiv:1511.05440, 2015.

[42] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv:1411.1784, 2014.

[43] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 14–29, Jan. 2016.

[44] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: A baseline for generative models of natural videos," arXiv:1412.6604, 2014.

[45] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3302–3309.

[46] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 98–106.

[47] N. Kalchbrenner, A. V. D. Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu, "Video pixel networks," arXiv:1610.00527, 2016.

[48] T. Xue, J. Wu, K. Bouman, and B. Freeman, "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 91–99.

[49] Y. Zhou and T. L. Berg, "Learning temporal transformations from time-lapse videos," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 262–277.

[50] J. Walker, C. Doersch, A. Gupta, and M. Hebert, "An uncertain future: Forecasting from static images using variational autoencoders," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 835–851.

[51] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," arXiv:1605.08104, 2016.

[52] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 8595–8598.

[53] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 2794–2802.

[54] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 527–544.

[55] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 737–744.

[56] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," *Adv. Neural Inf. Process. Syst.*, pp. 613–621, 2016.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980, 2014.

[58] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 314–327.

[59] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2929–2936.

[60] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1408–1424, Jul. 2015.

[61] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 921–928.

[62] 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 *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1153–1160.

[63] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" arXiv:1604.03605, 2016.

[64] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 233–240.

[65] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image Vis. Comput.*, vol. 30, no. 1, pp. 51–64, 2012.

[66] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 155–162.

[67] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, no. 10, pp. 1489–1506, 2000.

[68] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.

[69] T. Brox, C. Bregler, and J. Malik, "Large displacement optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 41–48.

[70] P.-H. Tseng, R. Carmi, I. G. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. Vis.*, vol. 9, no. 7, pp. 4–4, 2009.

[71] V. Buso, I. González-Díaz, and J. Benois-Pineau, "Goal-oriented top-down probabilistic visual attention model for recognition of manipulated objects in egocentric videos," *Signal Process.: Image Commun.*, vol. 39, pp. 418–431, 2015.

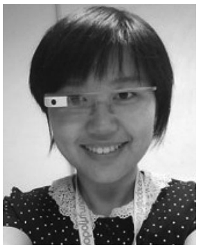
[72] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2714–2721.

[73] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 287–295.

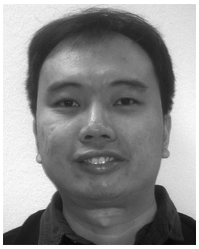
[74] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks,"in *Proc. IEEE Int. Conf. Comput. Vis.*, 2014, pp. 4489–4497.

[75] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2/3, pp. 107–123, 2005.

[76] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Visual Surveillance Perform. Eval. Tracking Surveillance*, 2005, pp. 65–72.



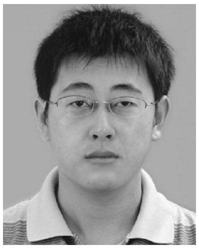
Mengmi Zhang received the BEng (first class honours) degree in electrical and computer engineering (ECE) from the National University of Singapore (NUS), Singapore, in 2015. She studied in University of California, Santa Barbara as an exchange student, in 2014. She is currently working toward the PhD degree in the Graduate School for Integrative Sciences and Engineering, NUS. Her research interests include computer vision, machine learning, and cognitive neuroscience. She is a student member of the IEEE.



Keng Teck Ma received the PhD degree from the National University of Singapore (NUS), Singapore, in 2014. He is currently a research scientist with Agency for Science, Technology and Research, Singapore. His research areas include computer vision and human-centric artificial intelligence. In particular, he is interested in visual attention, memory, and ego-centric vision. He is a member of the IEEE.

Joo Hwee Lim received the BSc(Hons I) and MSc degrees in computing from the National University of Singapore, Singapore, and the PhD degree in computer science and engineering from the University of New South Wales (UNSW), Australia. He is the principal scientist and head (visual intelligence), with the Institute for Infocomm Research, A*STAR, Singapore. He has published more than 260 international refereed journal and conference papers and co-authored 25 patents (awarded and pending) in computer vision and pattern recognition. He is a senior member of the IEEE.

Qi Zhao received the PhD degree in computer engineering from the University of California, Santa Cruz, in 2009. She is an assistant professor with the Department of Computer Science and Engineering, University of Minnesota (UM), Twin Cities. She did her postdoc with the Computation and Neural Systems, and Division of Biology, Caltech from 2009 to 2011. Prior to joining UM, she was an assistant professor in ECE and the Department of Ophthalmology, National University of Singapore, Singapore. Her research interests include computer vision, machine learning, cognitive neuroscience, and mental disorders. She is a member of the IEEE.



Jiashi Feng received the PhD degree from the National University of Singapore, in 2014. He is currently an assistant professor with the Department of Electrical and Computer Engineering, National University of Singapore. His research areas include computer vision and machine learning. In particular, he is interested in object recognition, detection, segmentation, deep learning, and robust learning. He is a member of the IEEE.