Patch-level Gaze Distribution Prediction for Gaze Following

Qiaomu Miao Minh Hoai Dimitris Samaras
Stony Brook University, Stony Brook, NY 11794, USA
{qiamiao, minhhoai, samaras}@cs.stonybrook.edu

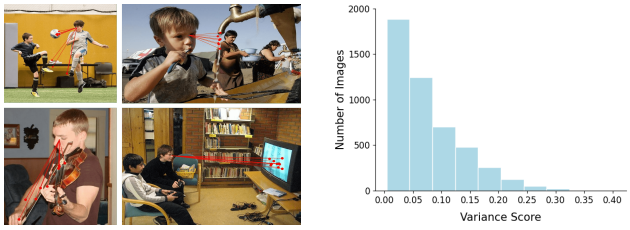
Abstract

Gaze following aims to predict where a person is looking in a scene, by predicting the target location, or indicating that the target is located outside the image. Recent works detect the gaze target by training a heatmap regression task with a pixel-wise mean-square error (MSE) loss, while formulating the in/out prediction task as a binary classification task. This training formulation puts a strict, pixel-level constraint in higher resolution on the single annotation available in training, and does not consider annotation variance and the correlation between the two subtasks. To address these issues, we introduce the patch distribution prediction (PDP) method. We replace the in/out prediction branch in previous models with the PDP branch, by predicting a patch-level gaze distribution that also considers the outside cases. Experiments show that our model regularizes the MSE loss by predicting better heatmap distributions on images with larger annotation variances, meanwhile bridging the gap between the target prediction and in/out prediction subtasks, showing a significant improvement in performance on both subtasks on public gaze following datasets.

1. Introduction

Gaze behavior is an important human behavior that serves a crucial role in inferring social intent and interactions [9, 1, 24], assisting human-computer interaction [21, 19], predicting learning outcomes [26], and diagnosis of psychological disorders like autism [4, 13, 3]. Therefore, analyzing human gaze automatically has attracted significant interest from computer vision researchers. Specifically, gaze following seeks to understand human gaze behavior in the wild by predicting the gaze target of a person inside a scene image in a third-person view, by locating the gaze target if it is located within the image, or indicating that the target is located outside.

Recent gaze following work formulated the target detection task as a heatmap prediction task [18, 7, 10, 31, 2]. The heatmap prediction module is typically trained with a mean square error (MSE) loss with the ground truth heatmap, which is generated by applying a Gaussian kernel around the gaze target pixel. However, current gaze following datasets only provide one annotated coordinate for each person in the training set, while as in Figure 1, the gaze target is usually ambiguous as different annotators may disagree on the exact location of the gaze target. Therefore, always requiring the model to regress to a unimodal Gaussian distribution in a higher resolution is not only a strict constraint in regression, but also biases the model towards predicting the same distribution pattern during inference, which lacks the consideration of annotator disagreement. Therefore, a regularization method is needed to relax the stricter constraint of the pixel-wise MSE loss, and should also encourage the model to predict more general distributions instead of a single Gaussian for uncertain images.



(a) (b)

Figure 1: Variance of gaze target annotations. (a) Annotations from multiple annotators on example images in the GazeFollow[27] test set. Each red dot is the annotation from one annotator. Note that annotators often disagree on the exact gaze target. (b) The distribution of annotation variance for images in the GazeFollow test set. Please refer to Sec. 4.4.2 for the calculation of the variance score.

In addition, besides estimating the target location, an effective gaze following model should also be capable of indicating if the target is located outside the image (in/out prediction). Previous work only trained the model with MSE loss and binary cross entropy (BCE) loss on the heatmap and in/out probability score predicted from two heads for the target prediction and in/out prediction sub-

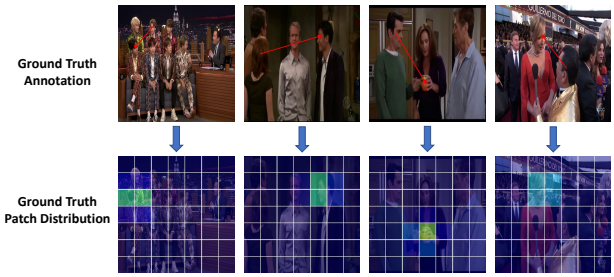


Figure 2: Visualizations of the ground truth patch distributions created from different images. Note that our method can create various patterns of distributions, compared to the Gaussian ground truth which is always circular.

tasks [6, 7, 10], without considering any correlation between them. However, we claim that these two subtasks should not be considered separately. Specifically, the outside case should be regarded as a special case of the target prediction task, except that the target is outside of the camera field-of-view. When a human follows someone's gaze in the image, he/she would directly perceive the location the person is looking at inside the image, or infer the person is looking at an unknown target outside the image, instead of considering the target and in/out prediction tasks separately. Therefore, we decide to model the gaze following behavior as a distribution of potential gaze targets over all possible locations, including an 'outside' target. This enables us to consider the two subtasks in a holistic manner, which better mimics human gaze following behavior, and brings significant improvement in the in/out prediction task.

In this paper, we propose the patch distribution prediction (PDP) for gaze following by replacing the in/out prediction task with the prediction of patch-level gaze distribution, which serves as a regularization for gaze heatmap prediction. Our PDP method regularizes the heatmap prediction from two perspectives. First, due to the coarser scale of patches, PDP has a softer constraint, which relaxes the pixel-wise MSE loss in higher resolution; In addition, as shown in Figure 2, our patch distribution (PD) has variable patterns for different images with our creation method. As the feature tokens are associated with the patch responses one-to-one, the variable distribution pattern and high responses in multiple patches will enhance the generality of the common feature tokens, and encourage the heatmap prediction head to predict multi-modal heatmaps in the coarser scale instead of a single Gaussian for more uncertain images. Furthermore, PDP also bridges the gap between the target prediction and in/out prediction subtasks by predicting gaze distribution. With the introduction of an 'outside' token, the gaze distribution can be predicted regardless of whether the target is located inside the image. Our claims are supported by our experiments.

Our main contributions can be summarized as follows:

- We propose the PDP method for gaze following. To the best of our knowledge, we are the first to address the imperfections of gaze following training methods by considering multiple scales, generalizing the target distribution patterns and the correlations between the in/out prediction and gaze target prediction subtasks.
- Our model is especially effective on images with larger variance in annotations. By predicting heatmaps that are more aligned with group-level human annotations, our model can achieve a much higher, super-human Area Under Curve (AUC) on the GazeFollow dataset.
- Our model also shows a significant improvement in the in/out prediction task. By integrating the target prediction and in/out prediction subtasks with PDP, our model enables training with two tasks simultaneously without loss in performance in either task.

2. Related Work

Gaze Following. GazeFollow [27] is the first work focusing on unconstrained gaze target prediction with a deep learning model. Its two-branch design, with one branch encoding the scene saliency information, and the other encoding the head gaze information has been commonly adopted in later works [6, 18, 7, 14]. Later, Recasens *et al.* [28] predicted the gaze target across temporal frames in videos. Chong *et al.* [6] considered the cases of the target located outside and trained the model in a multi-task learning approach. All these earlier methods formulated the gaze target prediction task as a one-hot patch classification task, which suffered from higher errors in distance metrics due to the coarse scale of patches.

Lian *et al* [18] was the first work that formulated gaze target prediction as a heatmap regression task, using the scene image and the multi-scale gaze direction fields. Chong *et al.* [7] proposed the VideoAtt model and extended the task to Video Gaze Following. Later, Fang *et al.* proposed the DualAtt model [10] by incorporating depth information, and 3D gaze direction estimated with eye images. Most recently, Tu *et al.* proposed a transformer model for gaze following [31], and Bao *et al* reconstructed the 3D scenes using depth maps and estimated human poses [2]. All these models trained the heatmap regression task with MSE, except for Lian et al [18] which uses binary cross entropy (BCE) loss. Despite some level of uncertainty consideration with Gaussian smoothing, the constraint to always regress to a circular Gaussian in a pixel-wise manner limits the model's generality for predicting distribution in uncertain cases. Furthermore, all previous models treated the in/our prediction task as a binary classification task with

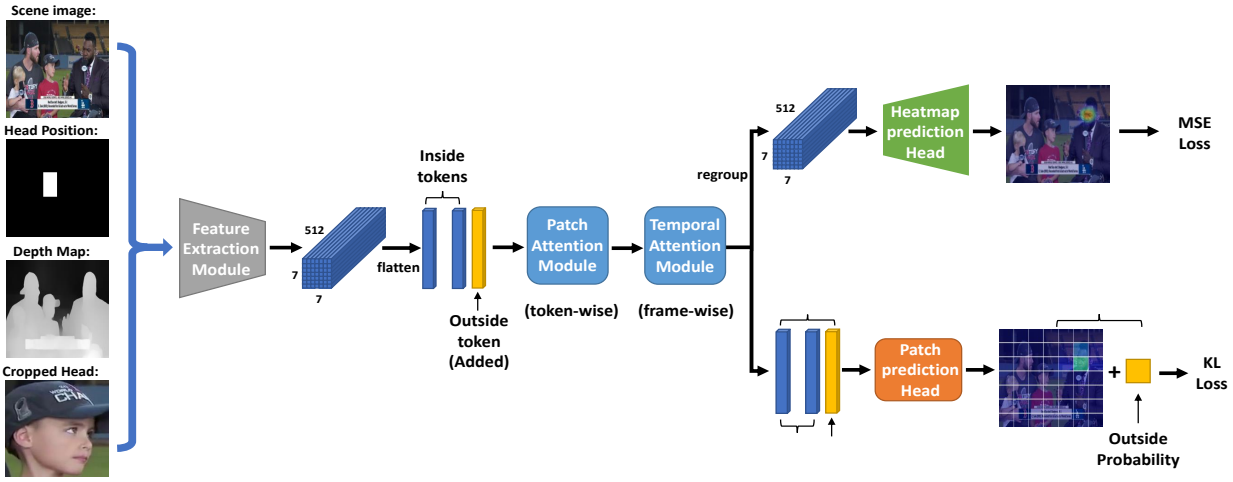


Figure 3: Overall structure of our model. The feature extraction module extracts feature encoding f_{enc} from the input, which is then flattened in the spatial dimension and concatenated with an added 'outside token'. The tokens go through the patch attention and temporal attention modules for information aggregation. Finally, the 'inside tokens' are regrouped in the spatial dimension for heatmap prediction, while all tokens go into another patch prediction head for PDP.

a separate head, while none has considered correlation between the two subtasks.

Perhaps the most similar work to ours is [34], which does a one-hot patch-level classification and a heatmap regression task for gaze prediction in egocentric videos. However, even if this method is optimal for egocentric gaze prediction, it is not optimal for the gaze following task due to the following reasons: The one-hot classification in the patch level is still unimodal with a fixed distribution pattern; The one-hot patch formulation may not represent the gaze distribution well if the target is close to the patch boundaries; The model cannot handle the case when the target is located outside. Furthermore, [34] shows cases of inconsistency between the patch classification and gaze prediction results, while our patch and heatmap predictions are highly consistent. We provide more detailed comparisons with the one-hot design in the supplementary material.

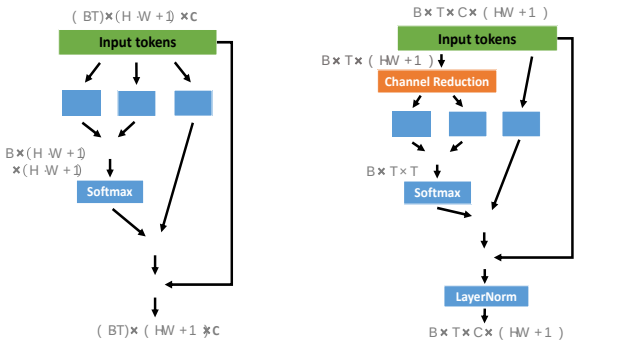
Gaze Object Prediction detects the gazed object or objects of interest that may attract gaze in a scene. Massé et al [20] predict the positions of all objects of interest that may be out-of-frame in a top-down view heatmap by inferring the gaze directions of people in a video, but requires multiple people in the scene and the object positions to be known in the top-down view during training. Recently, the GOO dataset [30] was released for gaze object prediction in retail environments, and a recent model was trained and evaluated on this dataset by doing object detection and target prediction with a shared backbone [32]. However, the GOO dataset contains no outside case and mostly synthetic images, and the ground truths are pre-determined and not annotated by human annotators.

3D Gaze Estimation predicts a 3D gaze direction instead of the gaze target. Methods of 3D gaze estimation can be divided into model-based [22, 29, 12, 35] and appearance-based approaches [33, 11, 5, 17, 23, 5, 15]. Model-based approaches estimate the gaze direction from geometric eye features and models. Appearance-based approaches estimate gaze direction directly from face or eye images. Some recent 3D gaze estimation methods consider the uncertainty caused by the varying levels of difficulties of the input. Kellnhofer *et al.* [15] used the pinball loss function [16] to compute the variance to a certain quantile for the gaze yaw and pitch angles. Dias *et al.* [8] used Bayesian neural networks to predict an uncertainty score with the input. However, these methods are difficult to be directly incorporated into gaze following tasks, as gaze following focus on target prediction instead of direction estimation, and the cases of multiple potential locations may make a single direction prediction suboptimal.

3. Method

3.1. Overall Structure

As shown in Figure 3, our model consists of three components: feature extraction, gaze distribution feature computation, and two heads for heatmap and patch-level gaze distribution prediction. The feature extraction module takes in the scene image $I \in R^{3\times H_0\times W_0}$, the binary head position mask $P \in \{0,1\}^{H_0\times W_0}$, a normalized depth map $D \in [0,1]^{H_0\times W_0}$, and the cropped head of the person $H \in R^{3\times H_0\times W_0}$ as input, and outputs the extracted feature $f_{enc} \in R^{C\times H\times W}$. The design of the feature extrac-



(a) (b)

Figure 4: Details of the (a) patch attention module and (b) temporal attention module.

tion module generally follows the VideoAtt model [7], except that we leveraged an additional depth map as input according to the insight from the DualAtt model [10] to incorporate scene depth information, with some small modifications. Please refer to the supplementary material for details of the feature extraction module.

Subsequently, the gaze distribution feature computation component operates on f_{enc} and outputs the gaze distribution feature $f_g \in R^{C\times(H\cdot W+1)}$, which consists of the inside tokens $f_g^{in} \in R^{C\times(H\cdot W)}$ and an outside token $f_g^{out} \in R^C$. Finally, f_g^{in} is regrouped and fed into the heatmap prediction head to output the heatmap $\hat{h} \in R^{C\times H'\times W'}$, while all tokens go into the patch prediction head, and output the patch-level gaze distribution $\mathcal{Q} = \{q_i\}_{i=1}^{H\cdot W+1}$. We will illustrate each component in detail below.

3.2. Gaze Distribution Feature Computation

This component is responsible for getting the gaze distribution feature before PDP and heatmap prediction. The output f_{enc} from the feature extraction module has a shape of $C \times H \times W$. We can consider that each C dimensional feature vector represent one patch in the image after dividing the image in to $H \times W$ patches. Therefore, f_{enc} can be regarded as $H \cdot W$ 'inside tokens' with C dimensions. f_{enc} is then flattened to the shape of $C \times (H \cdot W)$. To obtain the gaze distribution in both inside and outside cases, we add an 'outside token' $x_{out} \in R^C$, and concatenate it with the flattened and transposed version of f_{enc} and get $f'_{enc} \in R^{(H\cdot W+1)\times C}$. The 'outside token' is a learnable parameter vector that is trained along with the model. f_{enc} is then fed into the patch attention and temporal attention modules (Figure 4) to get the gaze distribution feature f_g.

The patch attention module $PA(\cdot)$ is introduced to make each patch token have a better understanding of the global scene information before computing the overall gaze distribution. It is a dot-product self attention module similar to [25], but we added learnable positional embeddings to all tokens to make the attention module aware of the positions of the tokens. In $PA(\cdot)$, each token in one spatial location is added with a weighted sum of all tokens, where the weights are computed from the dot product of the feature vectors in a pairwise manner and normalized by a softmax function:

$$\hat{f}_g = f'_{enc} + \text{softmax}(qk^\top)v, \tag{1}$$

where $q = W_q f'_{enc}$, $k = W_k f'_{enc}$, $v = W_v f'_{enc}$ are the queries, keys and values respectively. $W_q, W_k, W_v \in R^{C\times C}$ are learnable linear projections.

In video gaze following, the gaze distribution feature in nearby time steps should be helpful for the prediction of the current time step. To aggregate the information in the temporal dimension, we introduce the temporal attention module $TA(\cdot)$. The structure of $TA(\cdot)$ is similar to the patch attention module, but we made some modifications to make it work more efficiently. Suppose we have a sequence of output features $\hat{F}_g = \{\hat{f}_{gi}\}_{i=1}^T$ from a sequence of input images $\mathcal{I} = \{I_i\}_{i=1}^T$ in the video after the patch attention module. In order to reduce memory and computational load for calculating attention between frames, a convolutional layer is used to map $\hat{F}_g$ to 1 single channel and get $\hat{F}'_g \in R^{T\times(H\cdot W+1)}$ (we also tested mapping to more channels but did not observe improvement). Temporal attention weights are computed similarly as $PA(\cdot)$: $M_{att} = \text{softmax}(QK^\top)$, except the queries, keys and values are computed from the compressed feature map $Q = W_Q\hat{F}'_g$, $K = W_K\hat{F}'_g$, $V = W_V\hat{F}'_g$, $W_Q, W_K, W_V \in R^{(H\cdot W+1)\times(H\cdot W+1)}$. Finally, the original input features $\hat{F}_g$ are aggregated with the attention-weighted values with a residual connection to get the gaze distribution feature $F_g = \{f_{gi}\}_{i=1}^T$. We found applying a LayerNorm on the output can lead to more stable training:

$$F_g = LayerNorm(\hat{F}_g + f(M_{att}V)), \tag{2}$$

The temporal attention module will be removed for inference on a single image. In this case, the gaze distribution feature will be the output of $PA(\cdot)$: $f_g = \hat{f}_g$.

3.3. Gaze Prediction

The gaze distribution feature f_g is finally fed into two heads for gaze heatmap regression and PDP. The heatmap prediction head consists of one convolutional layer, followed by 3 deconvolutional layers, and a final convolutional layer, which is similar in structure to the ones in the VideoAtt [7] and DualAtt [10] model, but we replaced their in/out prediction head with our patch distribution prediction head, which consists of two fully connected layers operating on the channel dimension to get the patch probability score for each token:

$$\pi_q = \sigma(h_2(Relu(h_1(f_g)))), \tag{3}$$

where σ indicates sigmoid function. The PD can be obtained by normalizing each token's score:

$$q_i = q(g_i = 1|X) = \frac{\pi_q^i}{\sum_{j=1}^{H \cdot W+1} \pi_q^j}, \tag{4}$$

where q_i is the probability confidence of the gaze target locating in token i, X indicates all inputs to the model. Therefore, the probability that the target is located inside the image is the sum of scores from all inside tokens: $P_{in} = \sum_{j=1}^{H \cdot W} q_j$. For gaze heatmap regression, the $H \cdot W$ 'inside tokens' f_g^{in} are regrouped to a spatial feature map $\widetilde{f}_g^{in} \in R^{C' \times H \times W}$. $\widetilde{f}_g^{in}$ is then fed into the heatmap prediction module to generate the output gaze heatmap $\hat{h}$.

In training, the heatmap regression loss $\mathcal{L}_{hm}$ is the MSE loss between the predicted and ground truth heatmap. KL-divergence loss is chosen for the PDP task with the predicted and ground truth PD:

$$\mathcal{L}_{pd} = KL(q(g|X)||p(g|X)) \tag{5}$$

The final loss is a weighted sum of the two losses:

$$L = \lambda_1 \cdot \mathcal{L}_{hm} + \lambda_2 \cdot \mathcal{L}_{pd} \tag{6}$$

3.4. Ground Truth Patch Distribution Creation

In order to train the subtask of PDP, a ground truth patch-level gaze distribution is generated, of which the procedure is shown in Figure 5.

The ground truth patch distribution is created from the ground truth heatmap for the heatmap prediction task, which is generated by applying a gaussian kernel around the annotated gaze coordinate:

$$h(j,k) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(j-g_x)^2 + (k-g_y)^2}{2\sigma^2} \tag{7}$$

where $g = (g_x, g_y)$ is the annotated gaze coordinate.

For each patch, the points in the heatmap within that patch are located, and the maximum heatmap score is taken from these points as the probability score of that patch:

$$\pi_p^i = \max_{(j,k) \in \mathcal{N}(i)} (h(j,k)), \tag{8}$$

where $\mathcal{N}(i)$ is the pixels in the heatmap that are within patch i. We obtain the patch-level distribution value $\widetilde{\pi}_p^i$ by dividing π_p^i by the summed scores from all patches. If target is located outside, the outside token will have a probability of 1 and all the inside tokens will have a probability of 0:

$$p_i = \begin{cases} \widetilde{\pi}_p^i \quad if\ Y = 1\ else\ 0 & i \leqslant H \cdot W \\ 1 - Y, & i = H \cdot W + 1 \end{cases}, \tag{9}$$

$Y = 1$ if the target is located in the image, otherwise $Y = 0$.

Figure 5: Procedure of Ground truth PD creation.

With this discretization and normalization method, we can obtain various distribution patterns for the ground truth patch distribution, usually with high responses in multiple patches, as shown in Figure 2. By encouraging the model to predict higher responses in different patches, the model will tend to predict multi-modal heatmap clusters in the coarser scale on images with ambiguous targets, with a shared feature embedding before the two prediction heads. We provide the implementation details and results of other alternatives of the patch distribution creation settings in the supplementary material.

4. Experiments and Results

4.1. Datasets

GazeFollow [27] is a large-scale image dataset for gaze following. The dataset contains over 122K images in total with over 130K people inside the images, of which 4782 people were used for testing. For human consistency, 10 annotations were collected per person in the test set, while the training set just contain 1 annotation per person. Later, Chong *et al.* [6] extended it with more accurate annotations of whether the gaze target is located inside the image. **VideoAttentionTarget** [7] is a dataset for gaze following in videos. Video clips were collected from 50 different shows on Youtube, each of which has a length between 1-80 seconds. The dataset consists of 1331 head tracks with 164K frame-level bounding boxes, 109,574 in-frame gaze targets, and 54,967 out-of-frame gaze indicators. Both the training and test sets contain only 1 annotation per person.

4.2. Evaluation Metrics

Area Under Curve (AUC) is commonly adopted by the gaze following works [27, 6, 18, 7, 10] for assessing the confidence of the predicted heatmap. The ground truth in GazeFollow is the annotations from all 10 annotators, while in VideoAttentionTarget is the heatmap created from the single annotated coordinate. **Dist**: L^2 distance between annotated gaze coordinate and predicted location, determined as the point with maximum confidence on the heatmap. Specifically, in GazeFollow dataset, both **Min Dist.** and **Avg Dist.** are calculated. **In/Out AP**: Average Precision (AP) is used in the evaluation of In/Out prediction based on predicted probability of the gaze target locating in frame.

Method	Dep.	AUC ↑	Dist. ↓	
			Avg.	Min.
Random [27]	×	0.504	0.484	0.391
Center [27]	×	0.633	0.313	0.230
Fixed Bias [27]	×	0.674	0.306	0.219
GazeFollow [27]	×	0.878	0.190	0.113
Chong *et al.* [6]	×	0.896	0.187	0.112
Lian *et al.* [18]	×	0.906	0.145	0.081
VideoAtt [7]	×	0.921	0.137	0.077
VideoAtt_depth	✓	0.927	0.131	0.071
DualAtt [10]	✓	0.922	<u>0.124</u>	<u>0.067</u>
HGTTR [31]	×	0.917	0.133	0.069
ESCNet [2]	✓	<u>0.928</u>	0.126	/
Human	/	0.924	0.096	0.040
Ours w.o. dep.	×	<u>0.928</u>	0.131	0.072
Ours	✓	**0.934**	**0.123**	**0.065**

Table 1: Evaluation of the spatial model part on GazeFollow dataset. Best numbers are marked as bold and 2nd best are underlined. Dep. indicate depth input.

Method	Dep.	*In frame*		*Out of frame*
		AUC ↑	Dist ↓	AP ↑
Random [7]	×	0.505	0.458	0.621
Fixed Bias [7]	×	0.728	0.326	0.624
Chong *et al.* [6]	×	0.830	0.193	0.705
VideoAtt [7]	×	0.860	0.134	0.853
VideoAtt_depth	✓	<u>0.911</u>	0.118	0.861
DualAtt [10]	✓	0.905	**0.108**	<u>0.896</u>
HGTTR [31]	×	0.904	0.126	0.854
ESCNet [2]	✓	0.885	0.120	0.869
Human	/	0.921	0.051	0.925
Ours w.o. dep.	×	0.907	0.116	0.881
Ours	✓	**0.917**	<u>0.109</u>	**0.908**

Table 2: Evaluation of the full model on VideoAttentionTarget dataset.

4.3. Results

We evaluate the spatial part of our model (without temporal attention module) on the GazeFollow dataset [27]. Table 1 shows our model's performance with the current state-of-the-art (SOTA) models. For a fair comparison with VideoAtt [7], we modified its feature extraction module as ours, and train with the scene depth map as additional input. We name this model as VideoAtt_depth. It can be seen that our method outperforms the VideoAtt_depth model by a large margin in all metrics, and also outperforms the other newer models [10, 31, 2] in the AUC metric. The AUC is the most important metric that evaluates the model's predicted heatmap with the group-level annotations, and our model can achieve super-human AUC even without depth input. The smaller performance advantage in the distance metrics may be because the PDP task emphasizes on predicting the gaze distribution instead of a specific gaze coordinate, which is the maximum point in the heatmap. Besides depth input, the DualAtt model [10] takes cropped eye images as input using head pose and face keypoint estimation models, and the ESCNet [2] estimates human pose for 3D scene reconstruction. Our model does not require these additional inputs, and can achieve higher AUC than ESCNet which claims to have multi-modal output.

Table 2 summarizes the performance of our full model on the VideoAttentionTarget [7] dataset. Besides improvement in the target prediction metrics, our method shows a significantly higher in/out AP score compared to all SOTA models, closer to human performance. This supports our claim that the distribution prediction task bridges the gap between the target prediction and in/out prediction subtasks.

Figure 6 shows the qualitative results. It can be seen that our model can predict both single and multi-modal heatmaps and patch distributions. There are good correspondences between the predicted patch distributions, the predicted heatmaps and the gazed objects.

4.4. Analyses

4.4.1 Ablation Study

Table 3 summarizes our ablation study results. The model still has a strong performance without the temporal attention module. However, there is an obvious drop in performance after removing the patch attention module on VideoAttenionTarget. This is understandable because, without the patch attention module, the tokens lose some global context, and the outside token will have a fixed prediction score for any input, making the distribution prediction difficult. We then tested replacing KL divergence with MSE or BCE which were commonly used in gaze heatmap prediction [18, 6, 10, 31, 2]. Training with MSE showed a large drop in performance. BCE showed better performance than MSE due to higher robustness to noises, but is still worse than KL divergence. This demonstrates the importance of formulating PDP as a distribution prediction task. Finally, we replaced the PDP head with the original in/out prediction head in previous models [7, 10], by predicting an in/out probability score from f_g, and followed their training settings (no in/out loss for GazeFollow, and BCE for in/out loss in VideoAttentionTarget). The performance had a significant drop and became even worse than VideoAtt_depth. This validates the effectiveness of PDP, and also shows that simply introducing the 'outside token' and patch attention module without the supervision of patch distribution loss will hurt the performance. We also tried additionally predicting a 2D gaze direction from the head feature f_h, but

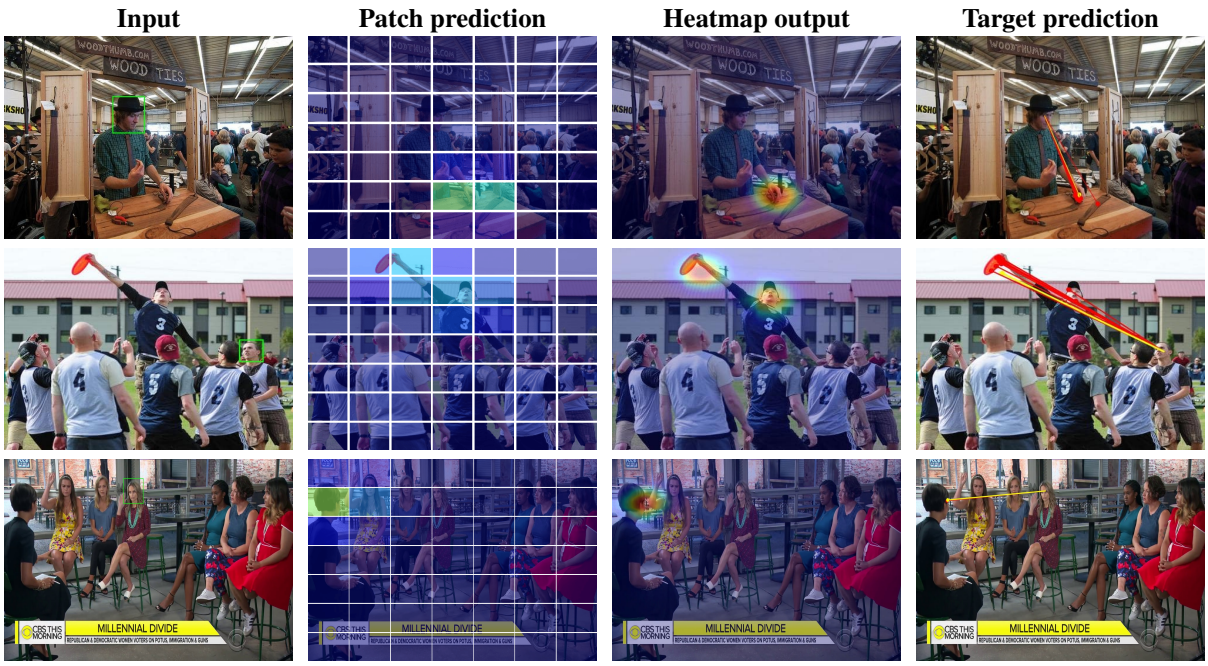


Figure 6: Visualizations of output of our model on GazeFollow (1st and 2nd rows) and VideoAttentionTarget dataset (3rd row). Ground truth annoatations are visualized in red and the predicted point is visualized in yellow.

Method	GazeFollow			VideoAttentionTarget		
	AUC ↑	Dist. ↓		*In frame*		*Out of frame*
		Avg.	Min.	AUC ↑	Dist ↓	AP ↑
No patch pred	0.927	0.138	0.077	0.889	0.130	0.885
KL → MSE	0.928	0.131	0.072	0.908	0.118	0.892
KL → BCE	0.931	0.127	0.067	0.904	0.116	0.890
No patch att	0.929	0.125	0.066	0.902	0.119	0.900
No temporal att	/	/	/	0.914	0.111	0.906
Add dir pred	0.931	0.124	**0.065**	/	/	/
Full Model	**0.934**	**0.123**	**0.065**	**0.917**	**0.109**	**0.908**

Table 3: Ablation Study Results

did not observe any improvement on GazeFollow.

4.4.2 Performance Regarding Annotation Variance

We evaluated our model's performance on images with larger annotation variance in the GazeFollow test set. To compute the annotation variance score for each image, we calculated the mean of the distances between each annotated coordinate and the average coordinate of all annotations. A larger distance indicates more disagreement between annotators. The statistics of the variance scores can be seen in Figure 1. We divided the dataset into 10 equal parts according to the quantiles of the variance score, each containing about 450 images. The model is evaluated in each part, both with and without depth maps as input, and compared with the corresponding version of the VideoAtt model. As we expect to examine the distribution of the heatmap predictions, we focus on the AUC metric.

Figure 7 shows our comparison results. It can be observed that our model shows an obviously higher AUC on images with variance scores above the 50% quantile, while the performances of our model and the VideoAtt model are highly close for images with variance scores below the 50% quantiles. To rule out the potential effect of other factors, we also computed the differences in performance between the VideoAtt_depth and VideoAtt model. Results show that the performance gain by incorporating depth maps as input

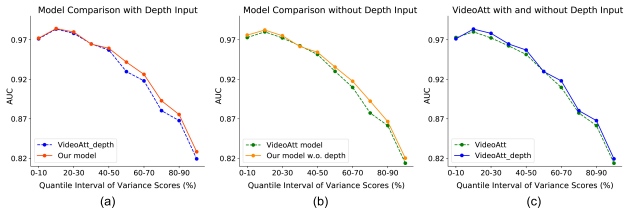


Figure 7: Comparison of AUC in each quantile interval in the GazeFollow test set. Our model shows obviously higher AUC in intervals with larger annotation variance, which cannot be observed in (c) when adding depth input only.

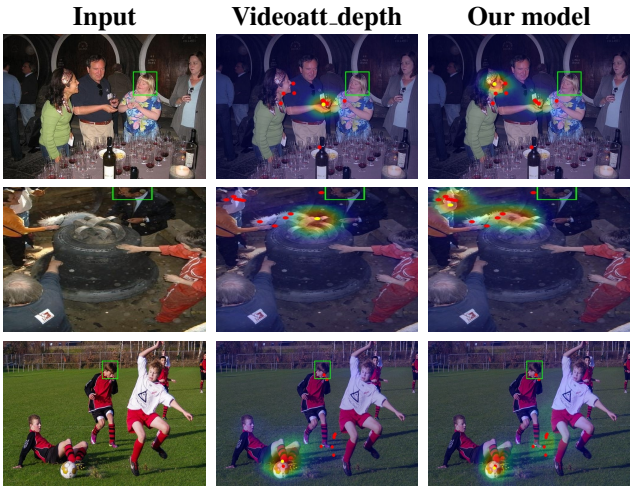


Figure 8: Predicted heatmaps from our model and VideoAtt_depth on images with larger "variance scores". The predicted (yellow) and ground truth (red) target points are also plotted. Our predicted heatmaps are more aligned with the group-level annotations.

lies evenly in the upper and lower quantiles, which in turn substantiates the claim that the performance gain from our method comes from the better heatmap predictions for images with larger variance in annotations by considering the uncertainty in gaze following annotations.

Figure 8 visualizes the predicted heatmaps of our model and the VideoAtt_depth Model on some example images with larger variance score. In contrast to the VideoAtt_depth model which only predicts a unimodal Gaussian, our model predicts multi-modal heatmap predictions which are better aligned with the group-level human annotations.

4.4.3 Comparison w/ Original In/Out Prediction Task

As mentioned earlier, when training on VideoAttentionTarget, the previous models [7, 10, 2] first pretrain the model on GazeFollow dataset, using MSE loss for the heatmap prediction task, and BCE loss for the in/out prediction task. However, as claimed in the official code of the VideoAtt model, in order to get SOTA performance on GazeFollow, the BCE loss was not involved in training. In our experiments, we found that when the model is trained with two losses together, there is a large drop in performance for the heatmap prediction task, as in Table 4. This seemingly weird result may stem from the separate handling of the two subtasks, causing introducing the in/out prediction task hurting the performance in the target prediction task.

Table 4: Effect of the In/out Prediction Task on Heatmap Prediction Task for VideoAtt model on GazeFollow dataset

Method	AUC ↑	Dist. ↓	
		Avg.	Min.
VideoAtt [7]	0.921	0.137	0.077
VideoAtt w. in/out	0.921	0.147	0.083
VideoAtt_depth	0.927	0.131	0.071
VideoAtt_depth w. in/out	0.927	0.145	0.083
Ours w.o. dep.	0.928	0.131	0.072
Ours	**0.934**	**0.123**	**0.065**

In contrast, our PDP method integrates the two subtasks without loss in performance, which also enables a much more efficient way of pretraining. Our model only needs to be trained once, instead of training two versions of the model to get the best performance on the GazeFollow dataset, and VideoAttentionTarget dataset respectively.

5. Conclusions

In this paper, we propose the PDP method in gaze following. By using the extracted feature vectors as inside tokens and adding an outside token, a patch-level gaze distribution is predicted. The PDP method can serve as a regularization method to the MSE loss for heatmap regression. Experiments show the superior performance of our method over the baseline models with an obviously better performance on images with larger annotation variance. Furthermore, the PDP task bridges the gap between the target prediction and in/out prediction tasks by showing a significantly higher AP, and provides a much simpler way of training gaze following models for any in-the-wild images/videos.

Acknowledgements. This project was partially supported by US National Science Foundation Awards IIS-1763981, IIS-2123920, DUE-2055406, the Partner University Fund, the SUNY2020 Infrastructure Transportation Security Center, and a gift from Adobe.

References

[1] Andrea Abele. Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior*, 10(2):83–101, 1986.

[2] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14126–14135, June 2022.

[3] Simon Baron-Cohen. *Mindblindness: An essay on autism and theory of mind*. MIT press, 1997.

[4] Tony Charman, John Swettenham, Simon Baron-Cohen, Antony Cox, Gillian Baird, and Auriol Drew. Infants with autism: an investigation of empathy, pretend play, joint attention, and imitation. *Developmental psychology*, 33(5):781, 1997.

[5] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 100–115, 2018.

[6] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398, 2018.

[7] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020.

[8] Philipe Ambrozio Dias, Damiano Malafronte, Henry Medeiros, and Francesca Odone. Gaze estimation for assisted living environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 290–299, 2020.

[9] Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000.

[10] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11390–11399, 2021.

[11] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018.

[12] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006.

[13] Vikram K Jaswal and Nameera Akhtar. Being versus appearing socially uninterested: Challenging assumptions about social motivation in autism. *Behavioral and Brain Sciences*, 42, 2019.

[14] Tianlei Jin, Zheyuan Lin, Shiqiang Zhu, Wen Wang, and Shunda Hu. Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.

[15] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019.

[16] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

[17] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[18] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018.

[19] Päivi Majaranta and Andreas Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, pages 39–65. Springer, 2014.

[20] Benoit Massé, Stéphane Lathuilière, Pablo Mesejo, and Radu Horaud. Extended gaze following: Detecting objects in videos beyond the camera field of view. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[21] Carlos H Morimoto and Marcio RM Mimica. Eye gaze tracking techniques for interactive applications. *Computer vision and image understanding*, 98(1):4–24, 2005.

[22] Atsushi Nakazawa and Christian Nitschke. Point of gaze estimation through corneal surface reflection in an active illumination environment. In *European Conference on Computer Vision*, pages 159–172. Springer, 2012.

[23] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019.

[24] Ulrich Pfeiffer, Leonhard Schilbach, Bert Timmermans, Mathis Jording, Gary Bente, and Kai Vogeley. Eyes on the mind: investigating the influence of gaze dynamics on the perception of others in real-time social interaction. *Frontiers in psychology*, 3:537, 2012.

[25] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[26] Sanjay Rebello, Minh Hoai, Yang Wang, Tianlong Zu, John Hutson, and Lester Loschky. Machine learning predicts responses to conceptual questions using eye movements. In *Proceedings of the Physics Education Research Conference (PERC)*, 2018.

[27] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[28] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the*

IEEE International Conference on Computer Vision, pages 1435–1443, 2017.

[29] Sheng-Wen Shih and Jin Liu. A novel approach to 3-d gaze tracking using stereo cameras. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):234–245, 2004.

[30] Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3125–3133, 2021.

[31] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2202–2210, June 2022.

[32] Binglu Wang, Tao Hu, Baoshan Li, Xiaojuan Chen, and Zhijie Zhang. Gatector: A unified framework for gaze object prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19588–19597, 2022.

[33] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.

[34] Zehua Zhang, David J Crandall, Chen Yu, and Sven Bambach. From coarse attention to fine-grained gaze: A two-stage 3d fully convolutional network for predicting eye gaze in first person video. In *BMVC*, page 295, 2018.

[35] Zhiwei Zhu and Qiang Ji. Eye gaze tracking under natural head movements. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 918–923. IEEE, 2005.