Looking here or there? Gaze Following in 360-Degree Images

Yunhao Li[1] Wei Shen[2*] Zhongpai Gao[2] Yucheng Zhu[1] Guangtao Zhai[1*] Guodong Guo[3]
[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University [3]Baidu
{lyhsjtu, wei.shen, gaozhongpai, zyc420, zhaiguangtao}@sjtu.edu.cn; guoguodong01@baidu.com

Abstract

Gaze following, i.e.*, detecting the gaze target of a human subject, in 2D images has become an active topic in computer vision. However, it usually suffers from the out of frame issue due to the limited field-of-view (FoV) of 2D images. In this paper, we introduce a novel task, gaze following in 360-degree images which provide an omnidirectional FoV and can alleviate the out of frame issue. We collect the first dataset, "GazeFollow360"[1], for this task, containing around 10,000 360-degree images with complex gaze behaviors under various scenes. Existing 2D gaze following methods suffer from performance degradation in 360-degree images since they may use the assumption that a gaze target is in the 2D gaze sight line. However, this assumption is no longer true for long-distance gaze behaviors in 360-degree images, due to the distortion brought by sphere-to-plane projection. To address this challenge, we propose a 3D sight line guided dual-pathway framework, to detect the gaze target within a local region (here) and from a distant region (there), parallelly. Specifically, the local region is obtained as a 2D cone-shaped field along the 2D projection of the sight line starting at the human subject's head position, and the distant region is obtained by searching along the sight line in 3D sphere space. Finally, the location of the gaze target is determined by fusing the estimations from both the local region and the distant region. Experimental results show that our method achieves significant improvements over previous 2D gaze following methods on our GazeFollow360 dataset.*

1. Introduction

Gaze behavior is an essential part of human behavior, which is significant in studying human social behavior, human-object interaction [23, 22, 10, 40, 27, 32, 11, 9, 44, 2, 13]. Gaze following [36], has been an active topic in the computer vision community, whose purpose is to predict the location where each human subject in a scene is looking at, given a 2D image containing one or more human subjects.

Rapid developments have been witnessed for gaze following methods [7, 26, 47, 8, 14], but they are restricted in 2D images or 2D videos, which easily suffer from the situation that gaze targets are out of frame, due to the limited field-of-view (FoV), as shown in Fig. 1(left). It is hard to perceive a whole surrounding scene in a 2D image. Unlike 2D images, 360-degree images capture the entire viewing sphere surrounding the optical center of a camera, which alleviates this issue. In addition, 360-degree images have gradually been utilized in various scenes. For instance, autonomous driving systems take 360-degree images as the input, and thus gaze following in 360-degree images can be used for human behavior understanding, such as human motion prediction, which can help detect the human attention to avoid traffic crash. Together with the fact that the prices of 360-degree cameras (e.g., Ricoh Theta S, Samsung Gear 360) have been reduced, it becomes promising to conduct gaze following research in 360-degree images.

In light of these facts, in this paper, we propose a new task: gaze following in 360-degree images. Compared with gaze following in 2D images, two challenges are encountered in this task: (1) Current gaze following approaches are deep learning based which are data driven, but there is no public available large dataset for gaze following in 360-degree images. (2) Previous 2D gaze following methods are built upon the assumption that a gaze target should be in the 2D sight line of the human subject in the 2D image plane coordinate, as shown in Fig. 1(middle), while this assumption is no longer true for long-distance gaze behaviors in 360-degree images, due to sphere-to-plane projection, as shown in Fig. 1(right).

To deal with the first challenge, we establish the first large scale dataset "GazeFollow360" for gaze following in 360-degree images by collecting 360-degree images from real world scenes, including various indoor and outdoor scenes. Our dataset contains around 10,000 high quality human-gazing target annotation pairs. Each gaze target lo-

*Corresponding author: Guangtao Zhai, Wei Shen
[1]The dataset is at https://michaelliyunhao.github.io/here-or-there

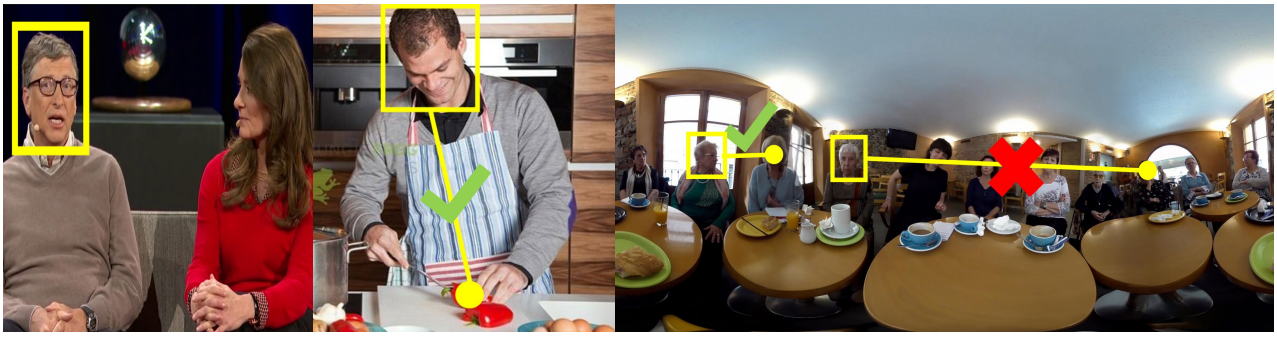

Figure 1. **Left**: The gaze target of the human subject in the 2D image is out-of-frame, due to the limited field-of-view of the 2D image. **Middle**: In a 2D image, the gaze target of a human subject is in his 2D sight line, since perspective projection preserves straight lines. **Right**: In a 360-degree image, this property still holds for short-distance gaze behaviors, while it is no longer true for long-distance gaze behaviors, due to the large distortion brought by sphere-to-plane projection. Thus our method copes with these two conditions parallelly by proposing a dual-pathway network.

cation is annotated by around 4 human labelers, and the final annotated location is the average. The dataset covers a wide range of potential application scenarios such as classrooms and sitting rooms, which can encourage development on gaze following in 360-degree images.

In addition, to address the second challenge, we propose a sight line guided dual-pathway framework for gaze following in 360-degree images. The second challenge is caused by the mismatch between the 2D sight line of a human subject and the gaze target to be looked at in 2D images, due to sphere-to-plane projection. The mismatch occurs when the human subject performs a long-distance gaze behavior, since the distortion brought by sphere-to-plane projection is large at this situation. While for a short-distance gaze behavior, the gaze target locates within a local region around the human subject's head. This local region on the sphere can be approximated by a plane, thus the distortion can be negligible and the assumption that the gaze target is in the 2D sight line still holds.

Based on these observations, we model the gaze sight line in 3D sphere space rather than in 2D image plane coordinate which avoids sphere-to-plane projection and reflects the propagation of sight line in real world more naturally. Guided by the predicted 3D gaze sight line, we propose a dual-pathway framework detects the gaze target within a local region (here) and from a distant region (there), parallelly. Concretely, given a human subject's head image, the direction of the sight line (gaze direction) is first estimated. Then, the local region is obtained as a 2D cone-shaped field along the gaze direction starting at the human subject's head position, and the distant region is obtained by searching along the gaze direction in 3D sphere space. Afterwards, gaze target estimation becomes attention guided saliency detection in both the local region and the distant region. The attention value at a position in the local region is inversely proportional to its angular difference to the sight line and that in the distant region is inversely proportional to its distance to the interaction between the sight line and the 3D sphere. Finally, the location of the gaze target is determined by fusing the estimations from both the local region and the distant region.

Our framework is inspired by the human perception process. To infer the gaze target of a human subject, humans used to first estimate a rough gaze direction of the human subject, then infer the possible regions of the gaze target, and finally confirm the location of the gaze target within the possible regions according to image content, such as object saliency.

The contributions of our paper are three-fold: **(1)** To our best knowledge, this is the first work that studies gaze following in 360-degree images. **(2)** We establish "GazeFollow360", the first large-scale dataset for gaze following in 360-degree images which contains 10,058 4K high resolution images with annotations of heads and gaze targets. **(3)** We propose a sight line guided dual-pathway framework to address the mismatch between the sight line of a human subject and the gaze target in 360-degree images.

2. Related Work

Gaze Following Some gaze following researches[20, 39, 30, 38, 1] paid attention to restricted scenes. [39] studied estimating the gaze target in a specific environment for human-robots interaction. Our work focuses on the general gaze following problem[36, 29, 8, 7, 28]. Recasens *et al.* [36] first defined the gaze following problem and built a 2D images dataset. Chong *et al.*[7] proposed a multi-task approach to learn gaze directions and saliency simultaneously. Further works also utilized the gaze directions to generate useful representation such as gaze fields [26, 8] or sight lines [47] to help gaze target prediction. Besides coping with 2D images, more works [8, 35, 53] extended the problem to videos or group gaze problem. However, these works are only concentrated on 2D images. Comparing with 2D images, there is a mismatch between the gaze target and the sight line of the human subject in 360-degree images, making typical 2D approaches suffer from perfor-

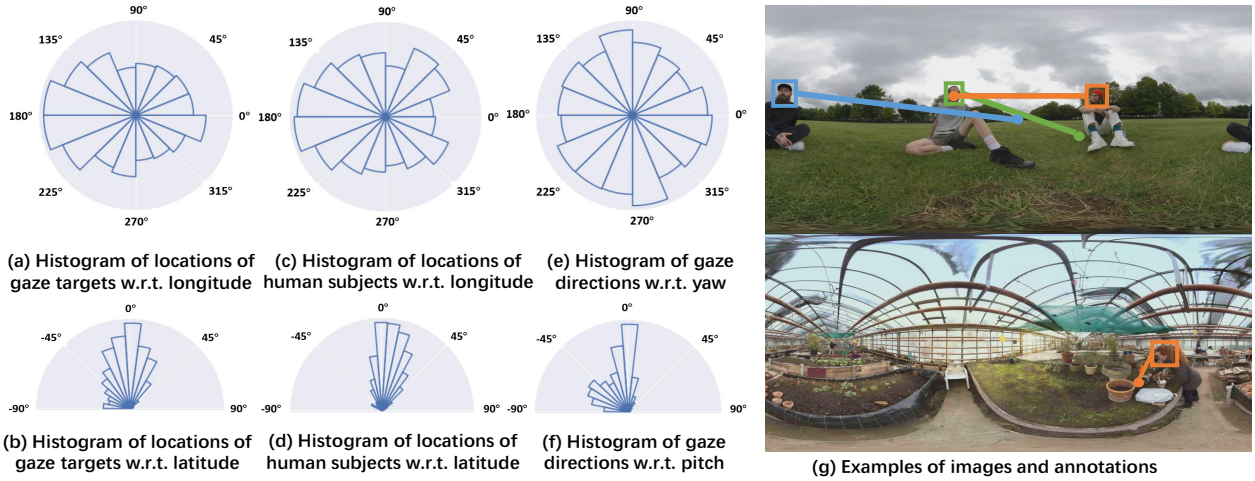

Figure 2. Overview of our GazeFollow360 dataset: (a) and (b) show the histograms of the locations of gaze targets w.r.t. longitude and latitude, respectively; (c) and (d) show the locations of human subjects w.r.t. longitude and latitude, respectively; (e) and (f) show the directions of gaze targets relative to the camera center w.r.t. pitch and yaw, respectively; (g) shows two examples of images and annotations.

mance degradation when dealing with 360-degree images.

Visual Saliency Prediction Visual saliency prediction is to estimate locations in an image which attract the attention of human subjects when looking at the image. Traditional saliency models are based on feature integration theory [41] and explore various hand-crafted features[50, 51, 5, 15, 18]. Recently deep learning based methods show superior performance on this task due to their strong ability of extracting features from images [24, 42, 48, 33, 4, 34, 43, 3, 25].

3D Gaze Estimation 3D Gaze estimation approaches can be categorized into model-based and appearance-based. Model based approaches[45, 17, 52] estimate gaze direction by constructing geometric eye models. Appearance-based approaches[46, 37, 21, 31, 6, 12, 49] seek to learn a mapping function from eye or head images to gaze directions. Recently, Zhang [46] and Cheng [6] utilize neural network to estimate gaze directions. Theoretically, 3D gaze estimation can directly find gaze targets in 3D sphere space, but it cannot achieve satisfactory gaze following results, as it ignores scene information in 360-degree images. We use it as the the first step of our framework, and refine its result by the following dual-path scene understanding modules.

3. The GazeFollow360 Dataset

We first construct the following large-scale GazeFollow360 dataset due to the lack of public available image/video dataset for gaze following in 360-degree images.

3.1. Data Collection and Annotation

In order to ensure that our dataset reflects the natural diversity of gaze behaviors, we collect 360-degree images from various scenes which are pretty common in the real world, such as sitting room and classroom. We concretely classify the scenes into indoor scenes and outdoor scenes. All these 360-degree images are crawled from YouTube videos. We select 65 different videos from YouTube covering various scenes, from which we extract short clips containing dynamic gaze behaviors. In each clip, we sample frames every 4 seconds, and totally we collect 10,058 frames, where each frame is a 360-degree image in the equirectangular format.

To annotate these 360-degree images for gaze following, we conduct an annotation pipeline as follows: First, the bounding boxes of the heads of the human subjects in each 360-degree image are labeled. Then, the gaze target of each human subject is labeled by 4 knowledgeable human annotators. Note that, even for human annotators, it is difficult to find gaze targets from equirectangular images directly, due to the distortion caused by sphere-to-plane projection. To tackle this difficulty, the annotators make use of a software "insta360 player", which can re-project an equirectangular image into 3D sphere space and view it from 360-degree, to find gaze targets.

We remove the annotated gaze target point which is obviously an outlier or noise. The mean of the rest of gaze target points is marked as the annotation result. This collection pipeline finally produces 10,058 gaze targets and 10,058 heads annotations. The gaze targets includes various of objects, such as faces, human bodies and man-made objects. We provide the histogram over the categories of the target objects in the supplementary material.

3.2. Dataset Statistics

Dataset Splitting We split the dataset we collected into training set, validation set and testing set, which contain 8225 images, 933 images and 900 images, respectively. Each set has both indoor and outdoor scenes. The whole dataset consists of training set (36 indoor scenes, 14 outdoor scenes), validation set (3 indoor scenes, 2 outdoor scenes) and testing set (7 indoor scenes, 3 outdoor scenes). There is no source-overlap among images from different sets.

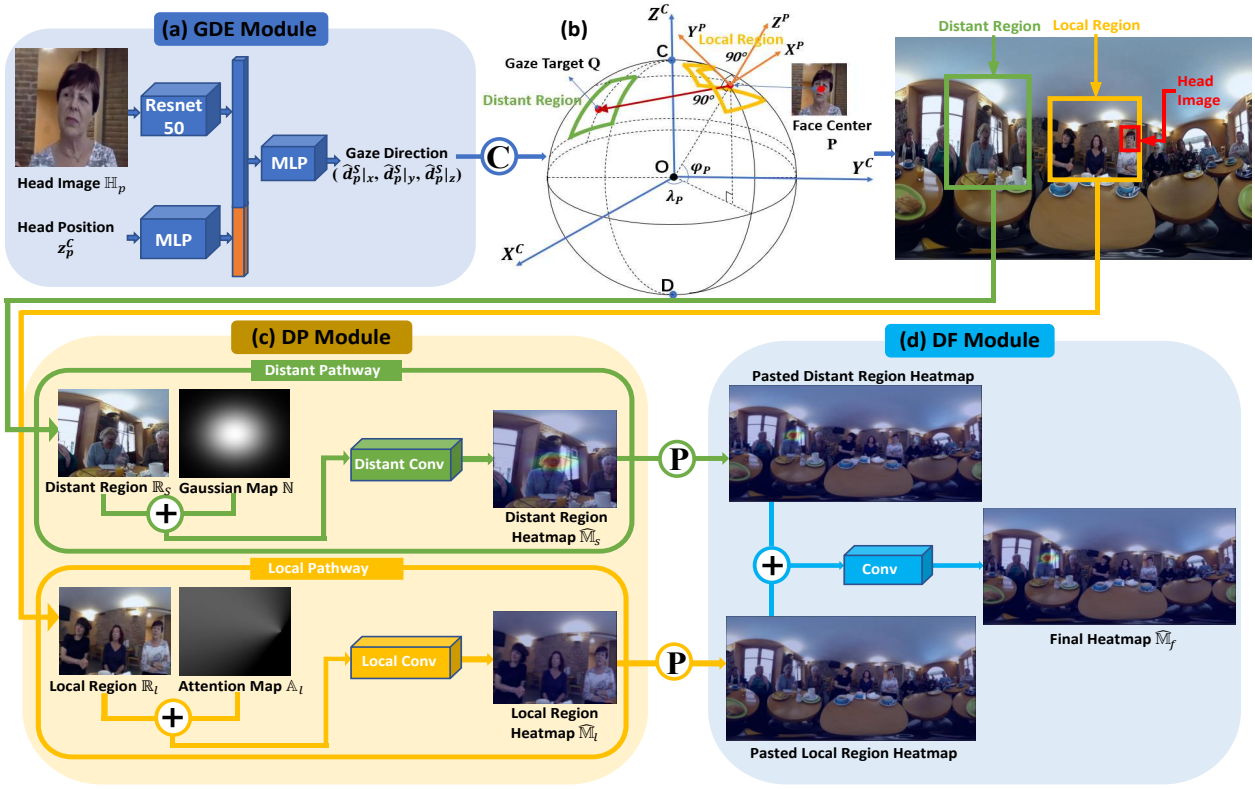

Figure 3. The overview of our framework. It consists of (a) a gaze direction estimation (GDE) module, (c) a dual-pathway prediction (DP) module and (d) a dual-pathway fusion (DF) module. "**+**" represents a concatenate operation. "**C**" represents a crop operation guided by the predicted gaze direction, which will be illustrated in section 4.4. "**P**" represents a paste operation which pastes the local predicted heatmap back to whole image. The local and distant crop regions are visualized in both the sphere (b) and the 360-degree image, marked by yellow and green, respectively. The example here is selected from the GazeFollow360 dataset, which shows a long-distance gaze behavior.

Annotation Statistics The annotation statistics of our dataset are shown in Fig. 2, including the histograms of gaze directions relative to the camera center w.r.t. pitch and yaw, and the histograms of the locations of gaze targets and human subjects w.r.t. longitude and latitude. The histograms of the locations of gaze targets and human subjects w.r.t. latitude show that gaze targets and human subjects tend to locate around the equator. The possible reason is that photographers tend to place 360-degree cameras at similar heights as gaze targets and human subjects. Our study also shows that the scatter of the locations of gaze targets along longitude is less severe than that along latitude, which almost covers all the longitudes uniformly. This indicates that 360-degree camera naturally captures the gaze targets and the human subjects along the whole longitude.

4. Method

4.1. Framework Overview

To deal with the mismatch between the sight line of a human subject and the gaze target to be looked at caused by sphere-to-plane projection, we propose a sight line guided dual-pathway framework. As shown in Fig. 3, the framework consists of three modules: 1) a gaze direction estimation (**GDE**) module, 2) a dual-pathway prediction (**DP**) module and 3) a dual-pathway fusion (**DF**) module.

First, the **GDE** module estimates the direction of the sight line (gaze direction) of a human subject in 3D sphere space rather than in the 2D image plane space, since 360-degree images are obtained by sphere-to-plane projection, *e.g.*, equirectangular projection. Then, by searching along the estimated gaze direction, the **DP** module finds two candidate regions within the local region of the human subject (here) and from a distant region to the human subject (there), respectively, which possibly contain the gaze target. Finally, the **DF** module combines the gaze estimations from the two pathways and provides the final predicted gaze target heatmap. In the rest of this section, we will introduce these modules in detail one-by-one.

4.2. Prerequisite

We first describe some prerequisite knowledge about coordinate transformations, including the transformation from the 2D image plane coordinate to the 3D sphere coordinate originated at camera (camera coordinate) and the transformation from the camera coordinate to the subject coordi-

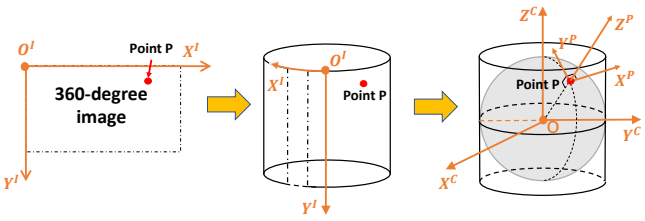

Figure 4. Image-to-sphere coordinate transformation and camera-to-subject coordinate transformation. (X^I, Y^I), (X^C, Y^C, Z^C) and (X^P, Y^P, Z^P) represent the image coordinate, the camera coordinate and the subject coordinate origined at P, respectively.

nate, as shown in Fig. 4. We use subscripts "I", "C" and "S", *e.g.*, $\mathbf{x}^I$, $\mathbf{x}^C$ and $\mathbf{x}^S$, to represent the coordinates in these three coordinate systems.

Image-to-sphere coordinate transformation: Let Q be a 2D point in a 360-degree image $\mathbb{I}$, *i.e.*, equirectangular image. Point Q is represented by $\mathbf{q}^I = (x_q^I, y_q^I)$, where $x_q^I \in [0,1]$, $y_q^I \in [0,1]$ are image plane coordinates. We can re-project Q back to a sphere by the reverse projection of equirectangular projection, which maps vertical straight lines of 2D image to meridians and horizontal straight lines of 2D image to circles of latitude. This re-projection process is shown in Fig. 4. In sphere space, we represent point Q in terms of sphere coordinates as $\tilde{\mathbf{q}} = (\varphi_q, \lambda_q, r) = (\pi/2 - \pi y_q^I, 2\pi - 2\pi x_q^I, r)$, where φ_q and λ_q represent the latitude and the longitude of point Q on the sphere, respectively, and r indicates the radius of the sphere. We then convert the sphere coordinates to the geocentric equatorial coordinates originated at the camera, and the representation of point Q becomes $\mathbf{q}^C = (x_q^C, y_q^C, z_q^C) = (r\cos\varphi_q \cos\lambda_q, r\cos\varphi_q \sin\lambda_q, r\sin\varphi_q)$.

Camera-to-subject coordinate transformation: Let P be the head center point of a human subject and Q be the corresponding gaze target point in a 360-degree image. We can obtain their representations in terms of camera coordinates $\mathbf{p}^C = (x_p^C, y_p^C, z_p^C)$ and $\mathbf{q}^C = (x_q^C, y_q^C, z_q^C)$ by applying the above mentioned image-to-sphere coordinate transformation. We can easily obtain the gaze direction by $\mathbf{d}_p^C = (x_q^C - x_p^C, y_q^C - y_p^C, z_q^C - z_p^C) = (d_p^C|_x, d_p^C|_y, d_p^C|_z)$. However, this formulation is not invariant to the rotation of the camera coordinate system, *i.e.*, the coordinate system originated at the camera O. To address this issue, we build a new coordinate system originated at the head center point of the human subject instead, *i.e.*, subject coordinate system, which is shown in Fig. 4 and Fig. 3(b). The Z-axis(Z^P) of the subject coordinate system is the extended line of line $\overline{OP}$; The Y-axis (Y^P) is orthogonal to line $\overline{OP}$ and also tangential to meridian CPD; The X-axis (X^P) is orthogonal to the plane formed by Y^P and Z^P. Intuitively, the gaze direction $\mathbf{d}_p^S = (d_p^S|_x, d_p^S|_y, d_p^S|_z)$ in terms of the subject coordinate system can be obtained by rotating the camera coordinate system.

$$\mathbf{d}_p^S = \mathbb{R}_p \mathbf{d}_p^C, \tag{1}$$

where $\mathbb{R}_p = \begin{pmatrix} -\sin\lambda_p & \cos\lambda_p & 0 \\ -\sin\varphi_p\cos\lambda_p & -\sin\varphi_p\sin\lambda_p & \cos\varphi_p \\ \cos\varphi_p\cos\lambda_p & \cos\varphi_p\sin\lambda_p & \sin\varphi_p \end{pmatrix}$ is the 3D rotation matrix. λ_p and φ_p are the longitude and latitude of the head center point P of the human subject on the sphere. One benefit to represent the gaze direction w.r.t subject coordinate system is it is independent to the subject's location, e.g., $\mathbf{d}_p^S = (0, 0, -1)$ when the subject looks directly at the camera.

4.3. Gaze Direction Estimation Module

Our purpose is to estimate the 3D gaze direction $\mathbf{d}_p^S$ of a human subject, given this subject's head crop image $\mathbb{H}_p$ and head center location $\mathbf{p}^I = (x_p^I, y_p^I)$ in a 360-degree image $\mathbb{I}$. Towards this end, we train a deep network $\mathcal{G}$ to regress the 3D gaze direction:

$$\hat{\mathbf{d}}_p^S = \mathcal{G}(\mathbb{H}_p, z_p^C), \tag{2}$$

where z_p^C is obtained by applying the image-to-sphere coordinate transformation introduced in Sec. 4.2 to $\mathbf{p}^I$. Here, we also use z_p^C as the input for gaze direction estimation, because z_p^C implicitly indicates the degree of distortion, *i.e.*, the bigger is the absolute value of z_p^C, the larger is the distortion at $\mathbf{p}^I$. Involving this distortion information into training can help our estimation module to be robust to distortion. The loss function for our network training is the cosine distance between the ground-truth gaze direction $\mathbf{d}_p^S$, which is computed by using the ground-truth gaze target location, and the estimated gaze direction $\hat{\mathbf{d}}_p^S$.

In our implementation, we feed the head crop image $\mathbb{H}_p$ into a ResNet-50 network [16] to extract facial features, as shown in Fig. 3(a). We then feed the head location z_p^C to a multi-layer perceptron (MLP) to obtain location features. Finally, we concatenate and feed them to another MLP for gaze direction estimation.

4.4. Dual-pathway Prediction Module

After estimating the 3D gaze direction $\hat{\mathbf{d}}_p^S$, we can use it to infer the candidate regions that contain the gaze target. One intuitive way is searching along the gaze direction in 3D sphere space, and finding the intersection between the sight line and the sphere. This strategy can address the mismatch between the sight lines and gaze targets in 360-degree images, especially for the long-distance gaze behaviors, which suffer from large distortion caused by sphere-to-plane projection. However, for the short-distance gaze behavior, this strategy becomes improper. The reason is a gaze behavior of a short distance implies the human subject is shown in profile in the 360-degree image, raising the difficulty to precisely estimate the 3D gaze direction in sphere space. Fortunately, the mismatch problem is not serious for short-distance gaze behaviors, since the distortion is small within a local range, and it can also be treated as common

2D gaze following problem. In light of this consideration, we design a dual path-way module to search the gaze target within a local region (here) and from a distant region (there), parallelly, according to the estimated gaze direction.

4.4.1 Distant Pathway

Given the estimated gaze direction $\hat{\mathbf{d}}_p^S = (d_p^S|_x, d_p^S|_y, d_p^S|_z)$ of a human subject as well as the location of head center $\tilde{\mathbf{p}} = (x_p^C, y_p^C, z_p^C)$, we can calculate the intersection point S between the estimated gaze sight line and the sphere by jointly solving the sphere equation and the gaze sight line equation in the camera coordinate system:

$$\begin{cases} (x_s^C)^2 + (y_s^C)^2 + (z_s^C)^2 = r^2 \\ \dfrac{x_s^C - x_p^C}{\hat{d}_p^C|_x} = \dfrac{y_s^C - y_p^C}{\hat{d}_p^C|_y} = \dfrac{z_s^C - z_p^C}{\hat{d}_p^C|_z} \end{cases}, \tag{3}$$

where $S^C = (x_s^C, y_s^C, z_s^C)$ is the location of the intersection point S and $\hat{\mathbf{d}}_p^C = (\hat{d}_p^C|_x, \hat{d}_p^C|_y, \hat{d}_p^C|_z)$ is the estimated gaze direction in terms of the camera coordinate system, *i.e.*, $\hat{\mathbf{d}}_p^C = \mathbb{R}_p^{-1}\hat{\mathbf{d}}_p^S$. Intuitively, if the estimated gaze direction is perfect, then the intersection point S should be the gaze target point. But, in practice, it is inevitable to have some errors in estimation. Thus, we crop a region $\mathbb{R}_s$ around the intersection point S from the 360-degree image $\mathbb{I}$ as the candidate region that contains the gaze target. Here, we define a cropping operation: $\texttt{Crop}[\cdot,\cdot,\cdot,\cdot]$, with four input parameters which are the whole image, the center of the crop, the width of the crop and the height of the crop, respectively. Then, we can express the crop region $\mathbb{R}_s$ by $\mathbb{R}_s = \texttt{Crop}[\mathbb{I}, \mathbf{s}^I, w_s, h_s]$, where $\mathbf{s}^I$ is the location of the intersection point S in image plane coordinate.

We then concatenate the crop region $\mathbb{R}_s$ with a Gaussian heatmap $\mathbb{N}$ of the same size whose peak is at the center of the region, and feed them into a deep network $\mathcal{G}_s$ to predict the heatmap $\hat{\mathbb{M}}_s$ of the gaze target within the crop region, as shown in Fig. 3(c):

$$\hat{\mathbb{M}}_s = \mathcal{G}_s(\mathbb{R}_s, \mathbb{N}). \tag{4}$$

The Gaussian heatmap $\mathbb{N}$ serves as an attention map to guide the predicted gaze target point to be close to the center of the crop region $\mathbb{R}_s$.

To train the network $\mathcal{G}_s$, we crop the ground-truth gaze target heatmap $\mathbb{M}$ corresponding to the crop region $\mathbb{R}_s$: $\mathbb{M}_s = \texttt{Crop}[\mathbb{M}, \mathbf{s}^I, w_s, h_s]$, and compute the BCE loss between $\mathbb{M}_s$ and $\hat{\mathbb{M}}_s$.

4.4.2 Local Pathway

For the short-distance gaze behaviors, since the distortion caused by sphere-to-plane projection is not large within a short range, we can search gaze targets along gaze directions directly in 2D image plane. Our basic idea is cropping a local region around the human subject's head and then producing an attention map with the same size of the cropped region guided by the subject's gaze direction. Our local Pathway is shown in Fig. 3(c).

Let $\hat{\mathbf{d}}_p^S = (\hat{d}_p^S|_x, \hat{d}_p^S|_y, \hat{d}_p^S|_z)$ be the estimated gaze direction of a human subject and $\mathbf{p}^C = (x_p^C, y_p^C, z_p^C)$ be the head center of this subject. One can compute the projected 2D gaze direction $\hat{\mathbf{d}}_p^I$ in the 2D image plane by the reverse process of the two coordinate transformations introduced in Sec. 4.2. Here, we apply a simple approximation strategy: In a short range, the projected 2D gaze direction $\hat{\mathbf{d}}_p^I$ in the 2D image plane can be approximated by $\hat{\mathbf{d}}_p^I \approx (-\hat{d}_p^S|_x, -\hat{d}_p^S|_y)$. We will give the detailed derivation in the supplementary material.

To obtain the local candidate region $\mathbb{R}_l$ around the head of the human subject, we adopt the following strategy: $\mathbb{R}_l = \texttt{Crop}[\mathbb{I}, \mathbf{p}^I - \text{sgn}(\hat{d}_p^S|_x) \cdot ((w_l - w_h)/2, 0), w_l, h_l]$, where $\text{sgn}(\cdot)$ is the sign function, $\mathbf{p}^I$ is the projection of $\mathbf{p}^C$ in the image plane, w_h is the width of the head of the human subject, and w_l and h_l are the width and height of the crop region, respectively. Here, we crop the region according to the horizontal component of the gaze direction, *i.e.*, $\hat{d}_p^S|_x$, since usually gaze targets are more distributed along the horizontal axis (longitude) than the vertical axis (latitude). We adjust w_l and h_l so that they are adaptive to the size of the head of the human subject because the size of the head implies the distance of the human subject to the camera and the scale of surrounding space. Hence, we set $w_l = \alpha_w w_h$ and $h_l = \alpha_h h_h$, w_h and h_h are the width and height of the head of the human subject, respectively, and α_w and α_h are hyper-parameters.

We then produce an attention map $\mathbb{A}_l$ with the same size of $\mathbb{R}_l$ guided by the projected 2D gaze direction $\hat{\mathbf{d}}_p^I$, which is a modified version of the gaze direction field (GDF) [26]. Let $\mathbf{p}^I = (x_p^I, y_p^I)$ be the projection of the head center point P in the image plane and M be an arbitrary point, with coordinates $\mathbf{m}^I = (x_m^I, y_m^I)$. GDF defines the probability that point M is the gaze point should be proportional to the angle between the direction $\mathbf{d}_{pm}^I = (x_m^I - x_p^I, y_m^I - y_p^I)$ of line $\overline{PM}$ and the projected 2D gaze direction $\hat{\mathbf{d}}_p^I$. Hence, the attention value at point M is computed by

$$\mathbb{A}_l(m) = \max(\left\langle \mathbf{d}_{pm}^I, \hat{\mathbf{d}}_p^I \right\rangle / \left(\left|\mathbf{d}_{pm}^I\right| \cdot \left|\hat{\mathbf{d}}_p^I\right| \right), 0). \tag{5}$$

Note that, the above computation does not take $d_p^S|_z$ into account, while the absolute value of $d_p^S|_z$ can imply whether the human subject is performing a short-distance gaze behavior or not. Recall that, a short-distance gaze behavior implies the human subject is shown in profile in the 360-degree image, which leads to a small absolute value of $d_p^S|_z$. Thus, we modulate Eq. (5) by introducing $d_p^S|_z$ into it:

$$\mathbb{A}_l(m) = \max(\frac{\left\langle \mathbf{d}_{pm}^I, \hat{\mathbf{d}}_p^I \right\rangle}{\left|\mathbf{d}_{pm}^I\right| \cdot \left|\hat{\mathbf{d}}_p^I\right|} \cdot \sqrt{1 - (\hat{d}_p^S|_z)^2}, 0). \tag{6}$$

We then concatenate the crop region $\mathbb{R}_l$ with the attention map $\mathbb{A}_l$ and feed them into a deep network G_l to predict the heatmap $\hat{\mathbb{M}}_l$ of the gaze target within the crop region.

$$\hat{\mathbb{M}}_l = \mathcal{G}_l(\mathbb{R}_l, \mathbb{A}_l). \tag{7}$$

To train the network $\mathcal{G}_l$, we crop the ground-truth gaze target heatmap $\mathbb{M}$ corresponding to the crop region $\mathbb{R}_l = \texttt{Crop}[\mathbb{I}, \mathbf{p}^I - \text{sgn}(\hat{d}_p^S|_x) \cdot ((w_l - w_h)/2, 0), w_l, h_l]$ and compute the BCE loss between $\mathbb{M}_l$ and $\hat{\mathbb{M}}_l$.

4.5. Dual-pathway Fusion Module

Since it is unknown that whether a human subject performs the a long-distance or short-distance gaze behavior, we combine the results provided by the local and the distant pathways to make the final prediction. We first create an all-zero map with the same size of the input 360-degree image, then we paste $\hat{\mathbb{M}}_l$ and $\hat{\mathbb{M}}_s$ back to this map and feed it into a 2 layer conv network $\mathcal{G}_f$ to generate the final heatmap $\mathbb{M}_f$ for the predicted gaze target, as shown in Fig. 3(d). The BCE loss is used to optimize $\mathbb{M}_f$ to approach the ground-truth gaze target heatmap $\mathbb{M}$. The predicted gaze target point is the peak location of the final heatmap.

5. Experimental Result

5.1. Experimental Setup

Dataset We evaluate our method and others on the new GazeFollow360 dataset. All the methods are trained on the training set, the optimal hyper-parameters for them are searched on the validation set, and the results on the testing set are reported for comparison.

Evaluation Protocol For evaluation metrics, we use the spherical distance and AUC in our experiments.

•*Spherical distance*: ℓ_2 distance is adopted as the evaluation metric for the previous 2D gaze following datasets. However, it is improper for our dataset, since a 360-degree image is the projection of a sphere. Thus, we use the spherical distance, a.k.a., great-circle distance, as our evaluation metric instead. It is the shortest distance between two points on the surface of a sphere. Let $\mathbf{q}^I = (x_q^I, y_q^I)$ and $\hat{\mathbf{q}}^I = (\hat{x}_q^I, \hat{y}_q^I)$ be the ground-truth gaze target point and the predicted gaze target point, assuming that they are projected on a unit sphere, then the spherical distance between them is

$$d_s = \arccos\left(\frac{<\hat{\mathbf{q}}^C, \mathbf{q}^C>}{|\hat{\mathbf{q}}^C| \cdot |\mathbf{q}^C|}\right), \tag{8}$$

where $\mathbf{q}^C = (\sin \pi y_q^I \cos 2\pi x_q^I, -\sin \pi y_q^I \sin 2\pi x_q^I, \cos \pi y_q^I)$ and $\hat{\mathbf{q}}^C = (\sin \pi \hat{y}_q^I \cos 2\pi \hat{x}_q^I, -\sin \pi \hat{y}_q^I \sin 2\pi \hat{x}_q^I, \cos \pi \hat{y}_q^I)$.

•*AUC*: Following [8], we use the Area Under Curve (AUC) criterion to assess a predicted gaze target heatmap.

5.2. Comparison to Other Methods

To validate the effectiveness of our framework, We compare our approach against following methods: (1) **random**: The predicted gaze target point is a randomly selected point in the 360-degree image. (2) **Free-viewing saliency prediction**: Free-viewing saliency prediction aims at identifying salient objects in images, which probably are gaze targets. We thus select a state-of-the-art saliency prediction method [19] as a baseline method for gaze following. (3) **3D gaze estimation**: 3D gaze estimation methods estimate the 3D gaze direction from head images. The gaze target point is the interaction point between the estimated 3D gaze slight line and the sphere, which can be located by searching along the estimated 3D gaze direction in sphere space. we compare our method against two typical 3D gaze estimation methods [46, 37]. (4) **2D images gaze following**: We select two recent 2D gaze following methods, [8] and [26], for comparison. [26] explicitly models the 2D gaze direction in the image plane coordinate and utilize it to help gaze target prediction. [8] is the current state-of-the-art 2D gaze following method which implicitly extracts gaze features. We compare with these approaches on GazeFollow360 dataset.

Method	Spherical Distance (↓)	AUC (↑)
Random	1.5357	0.5056
Lian *et al.* [26]	1.2540	0.6057
Salicon [19]	1.0940	0.8002
Chong *et al.* [8]	0.9183	0.7765
Ruiz *et al.* [37]	0.7612	0.7188
Zhang *et al.* [46]	0.7366	0.7213
GDE module	0.6880	0.7311
Ours	**0.6067**	**0.8104**
Human level	0.2531	0.9350

Table 1. Comparison results on GazeFollow360. "↓" and "↑" indicate the larger and the smaller the better, respectively.

GDE	DP	LP	Spherical Distance (↓)	AUC (↑)
✓			0.6880	0.7311
✓	✓		0.6410	0.7930
✓	✓	✓	**0.6067**	**0.8104**

Table 2. Ablation study of our framework on GazeFollow360. GDE, DP and LP represent the gaze direction estimation module, the distant pathway and the local pathway, respectively.

Comparison results are illustrated in Table 1. Our method outperforms others regarding the two evaluation metrics. Qualitative results are presented in Fig. 6, which shows the excellent ability of our framework.

It is worth noting that the 2D gaze following methods suffer from severe performance degradation. No matter whether explicitly predicting 2D gaze directions [26] or implicitly extracting gaze direction features [8] in the 2D image plane, they always encounter the severe mismatch problem that the gaze target is probably not along the 2D gaze direction in 360-degree images. The 3D gaze estimation methods [46, 37] achieve better performance than the 2D gaze following methods, since they model the sight line in

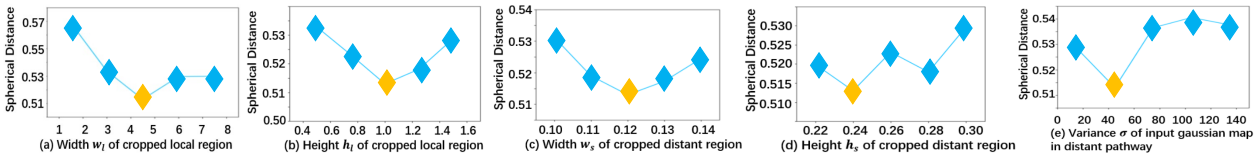

Figure 5. Sensitivity study for hyper-parameters on validation set: The width w_s and height h_s of the distant crop region, the width w_l, height h_l of the local crop region and the variance σ of the gaussian function used as the attention map in the distant pathway.

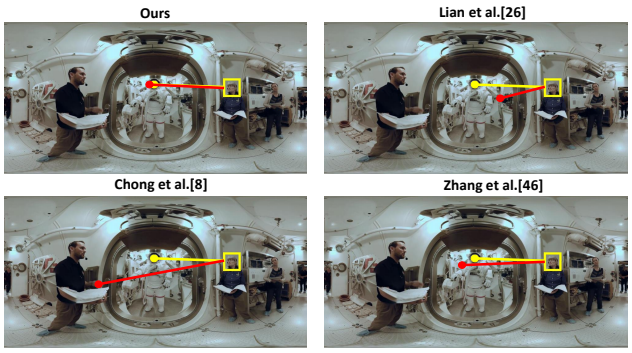

Figure 6. Qualitative results on GazeFollow360 dataset. The yellow and red points are ground truth and predicted targets, respectively. More results are provided in supplementary materials.

3D sphere space. But, they are worse than our GDE module, since our GDE module estimates 3D gaze directions by additionally encoding head positions. Our GDE module can be replaced by more advanced gaze estimation algorithms to achieve better results, but this is out of scope for our paper.

5.3. Detailed analysis

Ablation Study Now we conduct ablation experiments to verify the individual contribution of each module in our framework. The results are shown in Table. 2. We start by studying the gaze direction estimation (GDE) module. We use the same strategy for the 3D gaze estimation methods, as described in Sec. 5.2 , to locate the gaze target point for our GDE module. Comparing the result shown in Table. 2 and those in Table. 1, The GDE module already achieves a much better result than those of the 2D gaze estimation methods. This verifies our assumption that modeling gaze directions in sphere space is more efficient to deal with gaze following problem in 360-degree images.

Then, we observe that combining the GDE module and the distant pathway (DP) leads to a large improvement, since the DP performs a fine target search around the interaction point provided by the estimated 3D gaze direction.

Finally, we observe that further including the local pathway (LP), *i.e.*, our whole framework, the gaze direction guided dual-pathway framework, boosts the performance a lot and achieves the best result. This shows that the LP can re-detect the gaze targets missed by the DP. However, combining the GDE module solely with the LP cannot achieve satisfactory results, *i.e.*, 0.8747 spherical distance and 0.7142 AUC, which are even worse than the GDE module itself. The reason is the LP does not have the ability to detect long-distance gaze targets, as it only searches gaze targets within a local region at one side of the human subject, while the GDE module is able to locate both long-distance and short-distance gaze targets, as long as the estimated 3D gaze directions are correct. Note that, the GDE module only indicates at which side the local region is for the local pathway, rather than the candidate search region it delivers to the LP. Thus, the LP cannot benefit much from the combination with the GDE module. The ablation experiments show that 1) searching gaze targets in 3D sphere space is the key to gaze following in 360-degree images, 2) the local pathway is an important supplement to the distant pathway, and 3) combining these modules together can effectively detect both long- and short-distance gaze targets.

Sensitivity to Hyper-parameters We explore the sensitivity of the prediction results to the hyper-parameters involved in our framework, including the width w_s and height h_s of the distant crop region, the width w_l, height h_l of the local crop region and the variance σ of the gaussian function used as the attention map in the distant pathway. We evaluate the performance of each hyper-parameter on the validation set, which is shown in Fig. 5. It shows that our framework is robust to hyper-parameter changing in a certain range.

6. Conclusion

In this paper, we investigated a new task, gaze following in 360-degree images and collected a new large-scale dataset, "GazeFollow360", for this new task. We pointed out the main challenge of this new task is the mismatch between a human subject's gaze target and his/her sight line due to the distortion caused by sphere-to-plane projection in 360-degree images. To address this issue, we proposed a dual-pathway framework guided by sight lines modeled in 3D sphere space rather than simply in 2D image plane coordinate, to detect the gaze target within a local region (here) and from a distant region (there), parallelly. The strong performance of our framework on GazeFollow360 validates its potential for understanding gaze behavior in real 3D world.

Acknowledgments: This work was supported by NSFC 61831015, 61527804 and U1908210, Natural Science Foundation of Shanghai 21ZR1432200 and Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102.

References

[1] Arkar Min Aung, Anand Ramakrishnan, and Jacob R Whitehill. Who are they looking at? automatic eye gaze following for classroom observation video analysis. *International Educational Data Mining Society*, 2018. 2

[2] Sileye O Ba and Jean-Marc Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, 2010. 1

[3] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015. 3

[4] Ali Borji, Daniel Parks, and Laurent Itti. Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of vision*, 14(13):3–3, 2014. 3

[5] Neil DB Bruce and John K Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3):5–5, 2009. 3

[6] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10623–10630, 2020. 3

[7] 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)*, September 2018. 1, 2

[8] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 7

[9] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[10] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468, 2018. 1

[11] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5724–5733, 2019. 1

[12] 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. 3

[13] Kenneth A Funes Mora, Laurent Nguyen, Daniel Gatica-Perez, and Jean-Marc Odobez. A semi-automated system for accurate gaze coding in natural dyadic interactions. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 87–90, 2013. 1

[14] Jian Guan, Liming Yin, Jianguo Sun, Shuhan Qi, Xuan Wang, and Qing Liao. Enhanced gaze following via object detection and human pose estimation. In *International Conference on Multimedia Modeling*, pages 502–513. Springer, 2020. 1

[15] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[17] Craig Hennessey, Borna Noureddin, and Peter Lawrence. A single camera eye-gaze tracking system with free head motion. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, pages 87–94, 2006. 3

[18] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature: Highlighting sparse salient regions. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):194–201, 2011. 3

[19] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015. 7

[20] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 2

[21] 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. 3

[22] Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986. 1

[23] Michael Land and Benjamin Tatler. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press, 2009. 1

[24] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015. 3

[25] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3216–3223, 2013. 3

[26] 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. 1, 2, 6, 7

[27] Manuel Jesús Marin-Jimenez, Andrew Zisserman, Marcin Eichner, and Vittorio Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296, 2014. 1

[28] B. Massé, S. Ba, and R. Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2711–2724, 2018. 2

[29] B. Massé, S. Lathuilière, P. Mesejo, and R. 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, 2019. 2

[30] Zhixiong Nan, Tianmin Shu, Ran Gong, Shu Wang, Ping Wei, Song-Chun Zhu, and Nanning Zheng. Learning to infer human attention in daily activities. *Pattern Recognition*, page 107314, 2020. 2

[31] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. *arXiv preprint arXiv:1805.03064*, 2018. 3

[32] Cristina Palmero, Elsbeth A van Dam, Sergio Escalera, Mike Kelia, Guido F Lichtert, Lucas PJJ Noldus, Andrew J Spink, and Astrid van Wieringen. Automatic mutual gaze detection in face-to-face dyadic interaction videos. In *Proceedings of Measuring Behavior*, volume 1, page 2, 2018. 1

[33] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017. 3

[34] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 598–606, 2016. 3

[35] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[36] Adria Recasens Continente, Aditya Khosla, Carl Vondrick, and Antonio. Torralba. Where are they looking? In *In Advances in Neural Information Processing Systems*, 2015. 1, 2

[37] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018. 3, 7

[38] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Shor, Andrea Thomaz, and Scott Niekum. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8615–8621. IEEE, 2018. 2

[39] Akanksha Saran, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Understanding teacher gaze patterns for robot learning. In *Conference on Robot Learning*, pages 1247–1258, 2020. 2

[40] Omer Sumer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3327–3336, 2020. 1

[41] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 3

[42] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3183–3192, 2015. 3

[43] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3

[44] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6801–6809, 2018. 1

[45] Dong Hyun Yoo and Myung Jin Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding*, 98(1):25–51, 2005. 3

[46] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–60, 2017. 3, 7

[47] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *International Journal of Computer Vision*, pages 1–25, 2019. 1, 2

[48] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1265–1274, 2015. 3

[49] Wangjiang Zhu and Haoping Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3143–3152, 2017. 3

[50] Yucheng Zhu, Guangtao Zhai, and Xiongkuo Min. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69:15 – 25, 2018. 3

[51] Y. Zhu, G. Zhai, X. Min, and J. Zhou. The prediction of saliency map for head and eye movements in 360 degree images. *IEEE Transactions on Multimedia*, 22(9):2331–2344, 2020. 3

[52] 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. 3

[53] Ning Zhuang, Bingbing Ni, Yi Xu, Xiaokang Yang, Wenjun Zhang, Zefan Li, and Wen Gao. Muggle: Multi-stream group gaze learning and estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 2