ESCNet: Gaze Target Detection with the Understanding of 3D Scenes

Jun Bao[1] Buyu Liu[2] Jun Yu[1]*
[1]Hangzhou Dianzi University [2]NEC Laboratories America

Abstract

This paper aims to address the single image gaze target detection problem. Conventional methods either focus on 2D visual cues or exploit additional depth information in a very coarse manner. In this work, we propose to explicitly and effectively model 3D geometry under challenging scenario where only 2D annotations are available. We first obtain 3D point clouds of given scene with estimated depth and reference objects. Then we figure out the front-most points in all possible 3D directions of given person. These points are later leveraged in our ESCNet model. Specifically, ESCNet consists of geometry and scene parsing modules. The former produces an initial heatmap inferring the probability that each front-most point has been looking at according to estimated 3D gaze direction. And the latter further explores scene contextual cues to regulate detection results. We validate our idea on two publicly available dataset, GazeFollow and VideoAttentionTarget, and demonstrate the state-of-the-art performance. Our method also beats the human in terms of AUC on GazeFollow. Our code can be found here https://github.com/bjj9/ESCNet.

1. Introduction

Gaze target detection is important to understand human's intention. Therefore, it plays an important role in applications such as human computer interface [26] and social awareness tracking [27]. Though physical equipment such as wearable eye trackers [10] is available to perform gaze estimation, they are not desired due to location or calibration limitations. A more general setting takes third person view image as well as a given person in this scene as input and aims to locate where this person is looking in 2D image space [31]. Conventional methods typically leverage 2D visual cues to regulate gaze predictions by not only salient objects but also estimated gaze orientation [6, 31]. More recent approach [9] proposes to incorporate 3D gaze estimation and depth cues. Though demonstrating advanced performance,it requires additional human annotations [18] to explicitly model 3D gaze and due to coarse depth representation, it lacks the ability to handle more general scenarios, e.g. multiple salient objects lie in the same depth layer and field of view. To this end, we propose to perform gaze target detection by complete understanding and explicit modelling of 3D scenes, with 2D gaze annotations only.

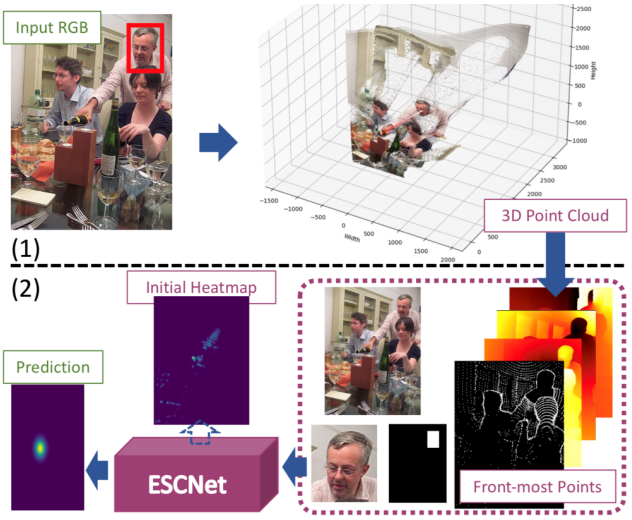

Figure 1. We propose to explicitly model 3D geometry in 2D gaze target detection task. We propose to reconstruct the scene with 3D point cloud in single-image setting in (1) and demonstrate that our ESCNet is able to effectively exploit such information in (2).

Promising as it sounds, lacking of 3D information in existing dataset makes our task hard. Also, effectively representing such information remains a vital problem.

We address these challenges through two key insights. Firstly, 3D geometry can be reconstructed by absolute depth and camera parameters, which can be estimated with relative depth and certain assumptions of reference objects. Specifically, we use "person" as our reference category as humans occur most frequently in images collected for gaze estimation task and their sizes are of certain distributions. With the assumption of human sizes, we can estimate the absolute depth and focal length of each image, leading to 3D point clouds (See (1) in Fig. 1). Second, occlusion plays an important role in gaze estimation given the fact that one cannot see through occluders. Such fact provides strong pri-

*Corresponding author.

ors to regulate where one person may look at. Inspired by this, we propose to represent the 3D geometry with front-most points, or occluders. This is achieved by modelling points in all possible 3D directions of a given person, and then leaving only the front-most one in each direction.

Further, we propose a novel model ESCNet that consists of gEometry and SCene parsing modules. The former leverages geometric cues, e.g. 3D gaze direction and 3D geometry, and outputs an initial heatmap as an intermediate representation, inferring the probability that each front-most point being looking at (See (2) in Fig. 1). The latter further incorporates scene contextual cues such as saliency in RGB image and generates final heatmap predictions.

We test on GazeFollow [31] and VideoAttentionTarget [6], and obtain state-of-the-art (SOTA) performance. Our intermediate representation is not only visually and conceptually meaningful, but also allows deep supervision, leading to performance boost. Finally, our method even outperforms human performance on AUC metric.

To summarize, our key contributions are:

- A novel method that explicitly models full 3D geometry, especially occlusion, in 2D gaze target detection.
- An end-to-end deeply supervised model ESCNet that explores 3D geometry, 2D/3D gaze and scene contextual cues, with gaze annotations only available in 2D.
- State-of-the-art results on publicly available datasets and superior performance over human.

2. Related Work

We organize our related work into three areas: gaze target prediction, 3D gaze estimation and 3D scene understanding from single image.

Gaze Target Detection Gaze target detection initially aims to locate gaze target of a given person in an image [4, 6, 9, 21, 31, 44]. The pioneer work [31] takes the first step towards gaze target detection and publishes a large-scale image dataset with annotations of head position and corresponding gaze targets. Following their design, most of the proceeding gaze target detection approaches [4,6,9,21] consider gaze and object saliency estimation in 2D image space when addressing this problem. Out of frame cases are first considered in [4, 6] where the person may look somewhere out of the image. One of the main limitations of existing work is that they rely on 2D visual cues and lacks the ability to reason in 3D. More recent method [9] proposes to incorporate depth cues to distinguish fore/background points. Impressive as it is, it requires additional 3D gaze labels [18] to specifically train its 3D gaze estimator. Without such 3D gaze labels, the core depth re-basing part of [9] will not work due to the lack of predictions in depth channel. In contrast, our work explicitly models 3D geometry with point clouds and effectively represents the 3D geometry by modelling only occluders from all 3D directions of given person, with gaze annotations only available in 2D. We further introduce an intermediate representation about the probability of the given person looking at each occluer. Such representation not only offers meaningful understanding of 3D scene and gaze, but also allows deep supervision.

3D Gaze Estimation 3D gaze estimation focuses on more basic gaze estimation problem where eye/face image of a single person is provided and its goal is to predict 3D gaze direction of this person. Existing methods can generally be categorized into model-based [2, 7, 15, 35] or appearance-based [25, 28, 33]. With the benefits of large scale datasets, e.g. MPIIGaze [43], CNN-based methods [1, 42, 43] further push gaze estimation field fast forward. Various techniques related to model design, including model structures, input and intermediate representations [29] have been explored. For instance, [11,43] propose complex or ensembles of CNNs to exploit their representation power and [3] models the two-eye asymmetry in face images. As for model input, multi-modal input [19, 39] and data normalization [41] have been proposed. 3D gaze also plays an important role in [9] to explicitly distinguish depth layers. In contrast to [9] that requires additional 3D gaze labels [18] to train its 3D gaze prediction module, we rely on gaze annotations in 2D only. Moreover, we explicitly model 3D geometry with front-most points and implicitly leverage 3D gaze cues in our model to regulate our predictions.

3D Understanding from Single Image Understanding 3D with only single image is an ill-posed problem as one single image can be generated from an infinite number of realistic scenarios [13]. Researchers have propose various representations for geometry, e.g., depth and normal [37], layer [16, 38] and layout [20, 23], semantics, e.g. 2D and 3D object [34], and combination of both [24, 40]. In this work, we choose point cloud as our representation as it is fine enough to model pixel-level occlusions. Specifically, we rely on predictions of relative depth from monocular method and that of reference objects to estimate the absolute depth and focal length. In order to model pixel-level occlusions, we represent each scene with occluders, or the front-most points in all 3D directions w.r.t. a given person.

3. Our Framework

As described above, our method consists of two stages and is illustrated in Fig. 2. The stage one explicitly models 3D geometry by reconstructing the entire scene from single RGB image, leading to 3D point clouds. At the second stage, our ESCNet effectively leverages the obtained 3D geometry to perform gaze estimation task. ESCNet consists of two sub-modules, the geometry and the scene parsing module. Specifically, the geometry parsing module (Sec. 3.1) estimates the probability of where a given person

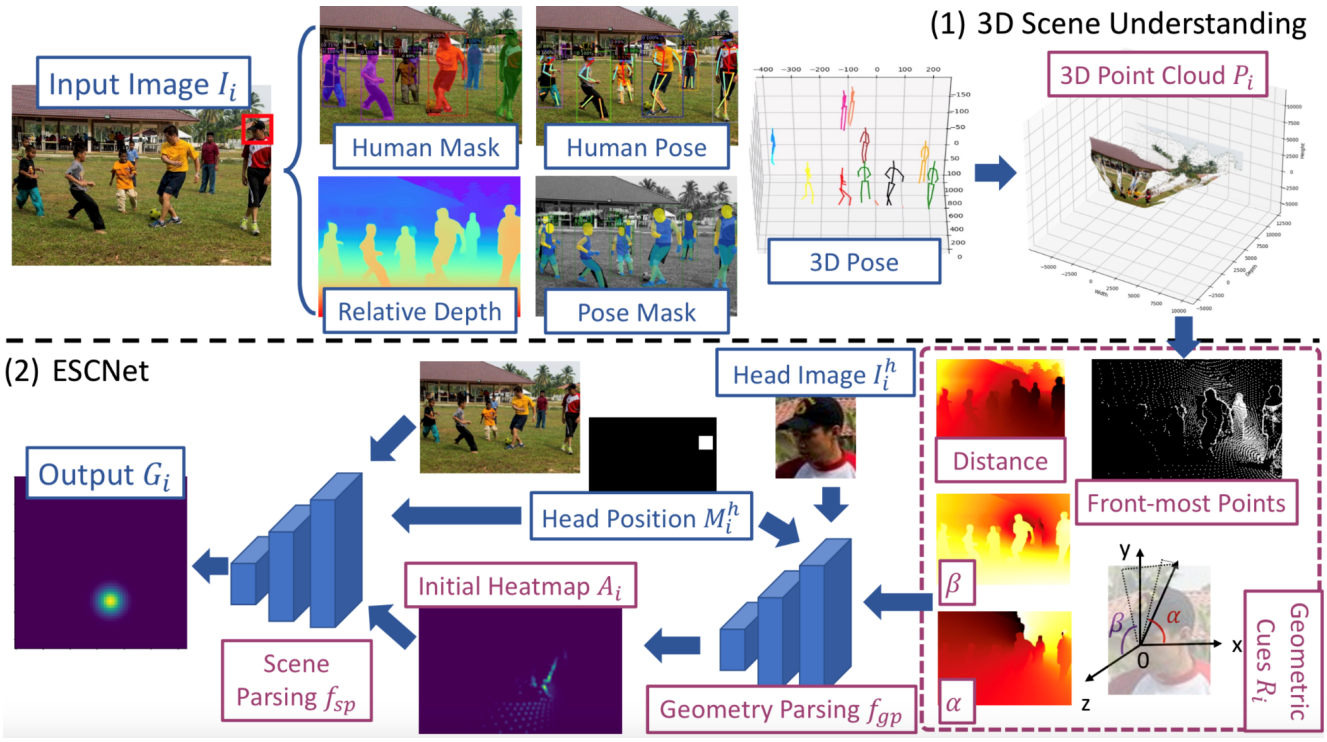

Figure 2. Our method takes a single RGB image and one given person as input and outputs the target gaze position in 2D. The stage (1) reconstructs the 3D scene with point clouds and in stage (2) ESCNet effectively represents the scene geometry with front-most points w.r.t. this person and makes predictions by further incorporating 3D gaze and scene contextual cues. Our ESCNet is deeply supervised and end-to-end trainable, with initial heatmap as an meaningful intermediate representation. We highlight our novelties in pink.

might look at by considering 3D geometric cues. The scene parsing module (Sec. 3.2) later incorporates both the scene contextual cues and the predicted probabilities to refine the target fixation prediction in 2D image space. For clarity, we first assume that 3D point clouds are available when introducing each sub-module and then describe how to obtain them from single RGB in Sec. 3.3.

Assuming that we have a dataset $\mathcal{D} = \{I_i, \mathbf{t}_i\}_{i=1}^N$ consists of N images as well as their annotations, where $I_i \in \mathbb{R}^{H_i \times W_i \times 3}$ denotes the i-th image, with height H_i and width W_i. $\mathbf{t}_i = [t_i^x, t_i^y]$ denotes the x, y locations of ground truth gaze fixation in 2D image space. We can automatically generate 3D point clouds of all images $\mathcal{P} = \{\mathcal{P}_i\}_i = \{\{\mathbf{p}_i^m\}_m\}_i$ and $\mathbf{p}_i^m = [p_i^{m,x}, p_i^{m,y}, p_i^{m,z}]$ is a 3-dimensional vector representing the 3D position of m-th pixel in the i-th image, with $m = [1, ..., H_i \times W_i]$. xy denote the image plane, and z is for depth direction. We further denote the gaze fixation and the head center in 3D space as $\mathbf{p}_i^t$ and $\mathbf{p}_i^h$. Equivalently, we can represent 3D points/vectors in angular space, or by its angle α, β and norm. Taking $\mathbf{p}_i^m$ as an example, $\mathbf{p}_i^m$ can be represented by $[f_\alpha(\mathbf{p}_i^m), f_\beta(\mathbf{p}_i^m), f_n(\mathbf{p}_i^m)]$, where $f_\alpha(\mathbf{p}_i^m) = \arctan 2(\frac{p_i^{m,y}}{p_i^{m,x}}) \in [-\pi, \pi]$, $f_\beta(\mathbf{p}_i^m) = \arccos(\frac{p_i^{m,z}}{\|\mathbf{p}_i^m\|}) \in [-\pi/2, \pi/2]$ and $f_n(\mathbf{p}_i^m) = \|\mathbf{p}_i^m\|$ denote the angle in coronal and sagittal plane, and norm of vector $\mathbf{p}_i^m$, respectively. The definition of angular space can be found in the bottom right of Fig. 2.

3.1. Geometry Parsing Module

The geometry parsing module f_{gp} aims to predict where a given person might view in 3D w.r.t. geometric cues.

The design of f_{gp} follows three main intuitions. Firstly, head image contains important information such as 3D head pose [46], which gives strong prior about gaze direction. Secondly, there exists strong correlations between where the head is located in 2D image and this person's fixation [31]. Lastly and most importantly, if there are multiple objects/points lie in one visual ray of one person, he/she can only focus on the closest object/point. Inspired by these three assumptions, f_{gp} takes head image $I_i^h \in \mathbb{R}^{H_i^h \times W_i^h \times 3}$, head position in 2D image $M_i^h \in \mathbb{R}^{H_i \times W_i}$ and the 3D geometric cues $R_i \in \mathbb{R}^{H_i \times W_i \times 4}$ as inputs. The output of f_{gp} is an initial heatmap $A_i \in \mathbb{R}^{H_i \times W_i} = \{a_i^m\}_m$ where a_i^m represents the probability of the given person in image I_i focusing on the m-th point in 3D space. We provide more details of f_{gp} in Fig. 3 and it is defined as:

$$A_i = f_{gp}(I_i^h, M_i^h, R_i) \tag{1}$$

Though it seems that our f_{gp} provides a probability map about where a given person might look at in the scene and is

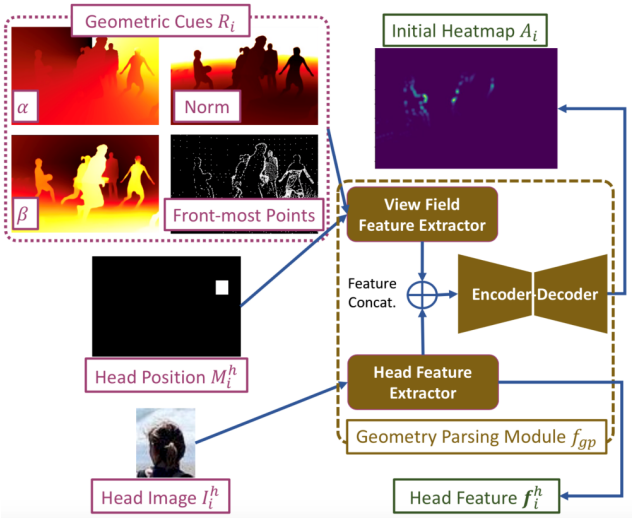

Figure 3. We provide more details of f_{gp}. Input, output and modules are visualized in pink, green and yellow, respectively.

similar to existing work [6, 9, 31], there are three main differences. Firstly, f_{gp} exploits the full 3D information while these existing methods either leverage only 2D cues [6, 31] or incorporate depth information at very coarse level [9]. Due to the 3D modelling, our intermediate representation A_i actually reflects probability of 3D points rather than pixels in 2D image space. Secondly, we explicitly represent front-most points in R_i and the intuition of R_i, or one cannot see through occluers so that only front-most points/objects are visible to given person, is lacking in existing methods. Finally, due to the explainable design of A_i, f_{gp} allows deep supervision for this intermediate representation during training while existing methods only receive heatmap-wise losses at the final prediction step. We will explain in below about how to obtain geometric cues R_i and ground truth A_i^*, and leave details of I_i^h and M_i^h to Sec. 3.3.

Geometric cues R_i consist of both the full 3D information of the current scene and the front-most point along each visual ray from the head center of a given person. Given the head center $\mathbf{p}_i^h$, the former can be easily obtain by converting all $\mathbf{v}_i^m = \mathbf{p}_i^m - \mathbf{p}_i^h$ to angular space, or $[\{f_\alpha(\mathbf{v}_i^m)\}_m, \{f_\beta(\mathbf{v}_i^m)\}_m, \{f_n(\mathbf{v}_i^m)\}_m]$. The latter is a binary map reflecting our intuition that when occlusion happens, only occluders are visible to human. In another word, if there are multiple objects/points along a visual ray, only the front-most one can be our focus. To achieve that, we first define visual rays and then figure out 3D points that lie in each visual ray. Lastly, we select out only the point with the minimum distance to a given person along each ray.

Given the head center $\mathbf{p}_i^h$, we assume that all visual rays must pass $\mathbf{p}_i^h$. Instead of working on the original continuous space, we propose to discretize them for better efficiency. Specifically, we discretize the coronal angle α into $J = 180$ bins and sagittal angle β into $K = 90$ bins, leading to 16200 possible discretized visual rays in total.

Figure 4. We visualize RGB image with the given person highlighted with red bounding box. The binary map of front-most points that this person can view is visualized on the right.

Our next step is to figure out which bin each point belongs to based on its angle w.r.t. head center. Denoting $\mathbf{v}_i^m = \mathbf{p}_i^m - \mathbf{p}_i^h, \forall m \neq h$, such angle can be represented by $[f_\alpha(\mathbf{v}_i^m), f_\beta(\mathbf{v}_i^m)]$. Based on the angle values and our discretization, we can further determine which bin each $\mathbf{p}_i^m$ falls into. When occlusion happens, there would be multiple points in one bin indicating more than one point occurs in this visual ray. We denote the $b_i^{m,\alpha} \in 1, ..., J$ and $b_i^{m,\beta} \in 1, ..., K$ as the bins that m-th point belongs to, e.g. it belongs to bin j, k if and only if $b_i^{m,\alpha} = j$ and $b_i^{m,\beta} = k$.

After determining the membership of each 3D point, we then remove occluded points in each bin. Or in another word, when multiple points belong to one bin, only the one with the smallest norm will be kept. We define a set $\mathcal{V}_i^{j,k}$ that consists of indexes of points that belong to the bin j, k. Then we have:

$$m_{j,k}^* = \arg\min_m f_n(\mathbf{v}_i^m), m \in \mathcal{V}_i^{j,k} \tag{2}$$

where $m_{j,k}^*$ denotes the index of point that has the minimal distance to head center in bin j, k. Grouping all such indexes together, or $\{m_{j,k}^*\}_{j=1,k=1}^{J,K}$, provides us the information about where the given person might look at in all possible 3D directions based on only 3D geometry. Then we generate a binary map of size $H_i \times W_i$ reflecting only the front-most points. If $m \in \{m_{j,k}^*\}_{j,k}$, we set its pixel value in this binary map to 1. Otherwise we set it to 0.

We provide some visual examples of the above mentioned binary map and the paired I_i in Fig. 4. We highlight the given person with red bounding box and show the generated binary map of front-most points. As can be seen in this figure, the generated binary map gives a satisfactory estimation about the front-most points in all possible 3D directions that a given person can focus on with scene geometry only.

Ground Truth A_i^* starts with the generated R_i and further incorporates gaze cues. Specifically, A_i^* aims to estimate the probability of each 3D direction that a given person is viewing. Intuitively, directions that are closer to ground truth gaze direction should have higher probabilities. Otherwise, their probabilities should be turned down.

Figure 5. We show four pairs of examples. We show on the left the RGB image, with the given person and the annotated ground truth highlighted. And our generated A_i^* is shown on the right.

To start with, we first generate the ground truth gaze direction, which can be obtained by $[f_\alpha(\mathbf{v}_i^t), f_\beta(\mathbf{v}_i^t)]$, where $\mathbf{v}_i^t = \mathbf{p}_i^t - \mathbf{p}_i^h$. We again discretize them into J and K bins by convolving a dirac delta function centered at $b_i^{t,\alpha}$, $b_i^{t,\beta}$ with a Gaussian of fixed variance, leading to the independent probabilities $\mathbf{Pr}(\alpha_i)$ and $\mathbf{Pr}(\beta_i)$. $Pr(\alpha_i)^j$ and $Pr(\beta_i)^k$ are the j-th and k-th value in vector $\mathbf{Pr}(\alpha_i)$ and $\mathbf{Pr}(\beta_i)$, denoting the probability of looking at the j-th and k-th discretized direction. Finally, we generate our probability map $A_i^* = \{\hat{a}_i^m\}_m \in \mathbb{R}^{H_i \times W_i}$ such that if $m \in \{m_{j,k}^*\}_{j,k}$, we have $\hat{a}_i^m = Pr(\alpha_i)^j \cdot Pr(\beta_i)^k$. Otherwise we set $\hat{a}_i^m$ to 0.

We visualize the generated A_i^* in Fig. 5. Again, we can see that A_i^* not only narrows down the target areas, but also provides meaningful probabilities w.r.t. ground truth 3D gaze directions compared to R_i. Our A_i^* takes into account the 3D cues thus it gives more diverse and meaningful guesses about where the given person might look at. One can expect that with contextual cues, we are able to refine our predictions one step further. Another interesting observation is about the multi-modal estimations and the potential ambiguity in human annotations. For instance, given only the single RGB image, it is hard to identify where the the girl in top-left figure really looks like at in pixel-level. She is more likely to focus on the left face or nose or mouth area of the lady than the right face/ear/shoulder due to its visibility. And we believe our A_i^* does reflect our observation. We will discuss this observation later in Sec. 4.

3.2. Scene Parsing Module

As discussed above, we have R_i that reflects where a given person might view in 3D space regardless of the head poses or contextual cues. We also obtain A_i that aims to predict the probability of each point/direction being looked at w.r.t. gaze related cues. Our next step is to incorporate contextual cues to refine our target position estimation.

Therefore, we introduce a scene parsing module that takes the current image I_i, the head position M_i^h, the probability map A_i and intermediate head features $\mathbf{f}_i^h$ from f_{gp} as input and outputs the final heatmap $G_i \in \mathbb{R}^{H_i \times W_i}$ reflecting the the confidence that a given person is fixating in each pixel location. Details of f_{gp} is shown in Fig. 6. Mathematically, we have:

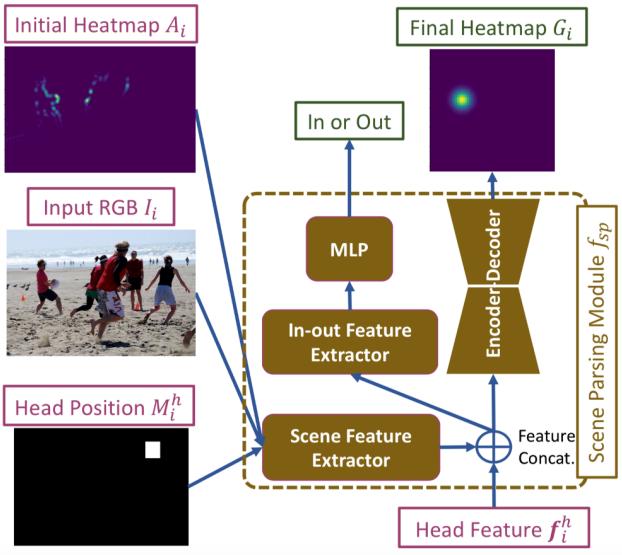

Figure 6. We provide more details of f_{sp}. Similarly, input, output and modules are visualized in pink, green and yellow, respectively. The heat feature $\mathbf{f}_i^h$ is obtained from f_{gp}.

$$G_i = f_{sp}(I_i, A_i, M_i^h, \mathbf{f}_i^h) \tag{3}$$

To obtain the ground truth G_i^*, we first generate a binary map where we set the value at $\mathbf{t}_i$ to 1 and elsewhere to 0. Then we convolve a dirac delta function centered at $\mathbf{t}_i$ with a 2D Gaussian of fixed variance.

Our overall loss function is then defined as:

$$\mathcal{L} = \mathcal{L}_{mse}(A_i, A_i^*) + \lambda \cdot \mathcal{L}_{mse}(G_i, G_i^*) \tag{4}$$

where $\mathcal{L}_{mse}$ denotes the MSE loss.

3.3. Data Preparation

In this section, we provide more details about how to obtain 3D point clouds $\mathcal{P}_i$, head image I_i^h and head position M_i^h with single RGB image.

To generate $\mathcal{P}_i$, we first estimate the relative depth of I_i and reference objects, or "persons". We rely on the assumption that human sizes are within certain distribution to estimate hyper-parameters. Specifically, the relative depth $D_i \in \mathbb{R}^{H_i \times W_i}$ is obtained by existing monocular depth estimator f_d by $D_i = f_d(I_i)$. $D_i^r = \frac{1}{a(D_i - b)}$ is absolute depth we desired. Our next step is to estimate D_i^r, or equivalently a, b, and focal length c for each image with the help from reference objects. Given the absolute depth map D_i^r as well as the focal length c, all pixels in I_i can be mapped to 3D with geometry [13], leading to our $\mathcal{P}_i$.

As described above, we use humans in I_i as reference objects. Specifically, we deploy pose estimator f_{dp} on each image and it gives masks of body parts of all persons. $E_i =$

$f_{dp}(I_i) \in \mathbb{R}^{H_i \times W_i \times C \times N_i}$, where C is the number of body parts and N_i denotes the number of individuals in the i-th image. We also exploit 2D/3D key point detectors, or f_{kd2} and f_{kd3} to provide detailed locations of body joints. Specifically, we have $K_i = f_{kd2}(I_i) \in \mathbb{R}^{C' \times 2 \times N_i}$ and $S_i = f_{kd3}(I_i)$, where $S_i \in \mathbb{R}^{C' \times 3 \times N_i}$ and C' denotes the number of key point categories.

For each individual l in I_i, we can obtain the tightest bounding box for head region from E_i. Then we denote the reciprocal of longest edge of this bounding box as $e_i^{h,l}$. We can further obtain its average relative depth on its head mask w.r.t. E_i and D_i, which we denote as $d_i^{h,l}$. Then b can be obtained by:

$$b = -\frac{\sum_l (e_i^{h,l} - \overline{e_i^{h,l}}) \cdot (e_i^{h,l} \cdot d_i^{h,l} - \overline{e_i^{h,l} \cdot d_i^{h,l}})}{\sum_l (e_i^{h,l} - \overline{e_i^{h,l}})^2} \tag{5}$$

where $\overline{*} = \frac{\sum_l (*)}{N_i}$ denotes the average function over $*$.

To compute a, we turn to the best represented person in image I_i. Intuitively, an individual that 1) provides large tightest mask 2) has majority of its key points detected in f_{kd2} and 3) gives high confidence score as person would be considered as our candidate. Denoting the index for the best represented person as $l*$, we can obtain the relative depth D_i over its detected full-body mask in E_i and denote it as $\mathbf{d}_i^{l*}$. We further get the maximum and minimum depth of this person by $\max(\mathbf{d}_i^{l*})$ and $\min(\mathbf{d}_i^{l*})$. Given the 3D key points of person $l*$, we can compute depth gap as well as width gap between any two key points. And we denote its maximum depth and width gap as $s_i^{d,l*}$ and $s_i^{w,l*}$. Then we can obtain a by:

$$a = \left(\frac{1}{\min(\mathbf{d}_i^{l*}) + b} - \frac{1}{\max(\mathbf{d}_i^{l*}) + b}\right) / s_i^{d,l*} \tag{6}$$

With a and b, we can easily obtain absolute depth with $D_i^r = \frac{1}{a(D_i - b)}$. Similarly, the average absolute depth of $l*$-th person is defined over its 2D mask and denoted as $d_i^{r,l*}$. By computing the tightest bounding box of this person in E_i, we can further get the width of this person, which we denote as $1/e_i^{w,l*}$. Then the focal length c is defined as:

$$c = d_i^{r,l*} / (s_i^{w,l*} \cdot e_i^{w,l*}) \tag{7}$$

We refer the readers to supplementary materials for more details to obtain a, b and c.

To obtain **head image** I_i^h, we directly exploit E_i to get the head mask. Then we crop image I_i w.r.t. the tightest bounding box of this mask to get I_i^h. Similarly, head **position mask** M_i^h is obtained by generating a binary map and set the values of only pixels that inside the above mentioned tightest bounding box to 1. To obtain **head center** $\mathbf{p}_i^h$, we use the key points of our target person detected in K_i, find the center of the key points for left and right eye location in 2D image space and then map this center to 3D. Our ground truth gaze target $\mathbf{p}_i^t$ is also obtained by mapping $\mathbf{t}_i$ to 3D. We refer the readers to Sec. 4 for more details.

4. Experiments

In this section, we demonstrate the effectiveness of our proposed method by conducting several experiments on two publicly available datasets, GazeFollow [31] and VideoAttentionTarget [6]. We demonstrate the state-of-the-art (SOTA) performances on these datasets and perform ablation study by validating the effectiveness of each module.

Datasets *GazeFollow* is a large scale gaze-folllowing dataset where 130,339 people in 122,143 images are collected from various existing datasets, e.g. ImageNet [8], with diverse activities and annotated with Amazon's Mechanical Turk (AMT). Following the split in [31], 4,782 people of GazeFollow are used for testing and the rest are for training. To ensure the annotation quality, every individual in the same image belongs to the same split and overall their fixations are uniformly distributed across the image. More importantly, to evaluate the human performance, 10 human annotations are collected per person on test images. *VideoAttentionTarget* gathers videos from 50 different shows from YouTube. And short clips ranges from 1 to 80 seconds are extracted from these shows where dynamic gaze behaviors as well as a person of interest can be continuously observed. During annotation process, both head bounding box and gaze target point of this person are annotated densely in clips, leading to 164,541 frame-level bounding boxes and corresponding gaze targets. About 20% of annotations are held out for testing, or 31,978 gaze annotations in 10 shows [1].

Evaluation Metrics We adopt four evaluation metrics [6, 9, 31] to evaluate the performance of gaze following methods. Area Under Curve (**AUC**) criteria [17] exploits the predicted heatmap as confidences to produce an ROC curve. We follow [6] to have a fair comparison. Specifically, ground truth is a binary map with the size of original RGB image where target locations from 10 annotations are set to 1 on GazeFollow dataset. While on VideoAttentionTarget dataset, the ground truth is obtained by thresholding a Gaussian confidence mask centered at the human annotator's target location. And AUC is measured under 64×64 resized space. L_2 Distance (**Dist.**) measures the Euclidean distance between the averaged ground truth annotations and predicted gaze location, or the averaged pixel location in the predicted heatmap. Note that height and width of images are all normalized to 1. Angular error (**Ang.**) reports the angular difference between the prediction and averaged

[1] Though the authors mentioned that 2 annotators work on test set [6], we only find one annotation available per test image in their released dataset. Thus we omit the comparison with human performance.

ground truth gaze vectors. Finally, Out of frame AP (**AP**) utilizes the average precision (AP) to assess the accuracy of out-of-frame identifying.

Implementation Details We use pre-trained MiDaS [30] as our monocular depth estimator f_d. As for pose estimator f_{dp}, we turn to Dense Pose [12] that pre-trained with COCO [22]. Our 2D key point detector f_{kd2} is of the same structure of X101-FPN [32, 36] that pre-trained on COCO [22]. We adopt pre-trained SMAP [45] as f_{kd3}. Since the original SMAP does not perform well on general gaze estimation dataset, we replace the 2D key points detection module in SMAP with the results obtained with f_{kd2}, or K_i, in practice. Note that we apply these pre-trained models directly on two datasets without re-training or fine-tuning. Our field view, head and scene feature extractors use ResNet50 as backbone [14]. The encoder-decoder in f_{gp}, f_{sp}, the in-out feature extractor and MLP share the same structure with that of [6,9]. We implement our method with PyTorch and use ADAM as optimizer and set the learning rate to 0.00025. Please note our method does not depend on the specific details of these sub-modules and we choose the above mentioned structures mainly for re-production purpose. We refer the readers to supplementary materials for more details about model structure of each components.

4.1. Performance on GazeFollow

Quantitative Results We demonstrate our quantitative results on GazeFollow dataset in Tab. 1. We highlight the best and second best number in bold and underline. To have a fair comparison, Video* does not include the temporal part of [6]. Ours+ replaces the head feature extractor in Ours with a model pre-trained on [18], or Whenet [46], so that the supervisions are the same w.r.t. [9]. **AP** is not reported as all annotations are in frame.

As can be seen from this table, compared to SOTA methods that require the same supervision, our method can always achieve superior performance. Even compared to method [9] that requires additional training data [18], Ours is comparable as well. We would like to highlight that our proposed method can beat the human performance under **AUC** metric. Unlike **Dist.** or **Ang.** that focuses on the average location of human annotations, which may be meaningless to some extend (See ground truth in Fig. 9), **AUC** actually provides a way to the measure the multi-modality property in predictions. Outperforming human performance shows that our method can indeed provides meaningful multi-modal and potentially more concentrated predictions on given images.

Visualization We visualize our results on GazeFollow dataset in Fig. 7. We demonstrate our prediction in yellow and the ground truth annotation in red. There are actually 10 annotations for each person during test time, we only visualize the average position of these 10 annotations in RGB

Method	Supervision [31]	Supervision 3D gaze	Evaluation Metric AUC↑	Dist.↓	Ang.(°)↓
Random [31]	✓		.504	.484	69.0
Center [31]	✓		.633	.313	49.0
Fixed bias [31]	✓		.674	.306	48.0
Recasens [31]	✓		.878	.190	24.0
Chong [5]	✓		.896	.187	-
Lian [21]	✓		.906	.145	17.6
Video* [6]	✓		.921	.137	-
Fang [9]	✓	✓	<u>.922</u>	<u>.124</u>	<u>14.9</u>
Human			.924	.096	11.0
Ours	✓		**.928**	.126	15.3
Ours+	✓	✓	**.928**	**.122**	**14.6**

Table 1. Evaluation on the GazeFollow dataset [31] for single-image gaze target detection. Numbers of baselines are from [6, 9].

Figure 7. We show our prediction and average human annotation in yellow and red, respectively.

and leave the discussion of the annotation reliability to later paragraphs. We show good examples on the left column while visualize our results in the middle and right column when human annotations are ambiguous. We can see that we can almost always give reasonable predictions.

Ablation Studies To evaluate the effectiveness of our 3D representation as well as intermediate representation, we perform two ablation studies. Firstly, we remove R_i in f_{gp} and train our model with only $\mathcal{L}_{mse}(G_i, G_i^*)$. In our second setting, we keep R_i but do not enforce losses in A_i. In both settings, the model structure the same as the ESCNet. Denoting the former as -geo-Aloss and the later as -Aloss, we report the performance on test set of [31] in Tab. 2.

By comparing -Aloss to our full model, we find that the absence of explicit modelling of A_i leads to inferior performance, which demonstrates the effectiveness of our intermediate representation and deep supervision. The performance drop from -Aloss to -geo-Aloss further showcases that R_i is truly effective and beneficial as 3D geometry rep-

Method	Evaluation Metric		
	AUC ↑	Dist. ↓	Ang. (°)↓
Ours	.928	.126	15.3
-Aloss	.921	.139	17.6
-geo-Aloss	.910	.161	21.1

Table 2. Evaluation on the GazeFollow dataset [31] for single-image gaze target detection. We gradually remove the loss for intermediate representation A_i and R_i to demonstrate the effectiveness of intermediate and our 3D geometry representation.

Figure 8. We visualize the RGB with given person on the left and our paired predicted A_i on the right. We can see that A_i reflects the 3D geometry and probabilities well.

resentation. We further visualize our predicted A_i in Fig. 8. We observe that our model can indeed generate meaningful intermediate representation about where a given person might look at w.r.t. 3D geometry. For instance, though clearly exists as salient object, the lady on the top-left cannot see the face of the kid in front of her due to occlusions. Similarly, the given person on the top-right figure cannot see the lady in red as she is occluded by the player in white.

Multi-modality in Human Annotation and Our Predictions Fig. 9 provides more details about our step-wise predictions and overall human annotations. From left to right, we visualize the input RGB image, our intermediate representation A_i, 10-annotation ground truths and our final prediction G_i. We can see that A_i and G_i share the same multi-modality property with real-world human annotations. Though missing in literature, we argue that such property is desired for gaze estimation task.

4.2. Performance on VideoAttentionTarget

We demonstrate our quantitative results on VideoAttentionTarget dataset in Tab. 3. Ours* shows the performance of directly applying our model that trained in GazeFollow to VideoAttentionTarget. To obtain **AP**, we just add additional BCELoss to Equ. 4. Again, we can always beat the SOTA approaches [5, 6] that require the same gaze supervision. Our performance is even better than Video [6], which is obtained with additional temporal cues. While ours is slightly worse than [9] that requires additional gaze-related

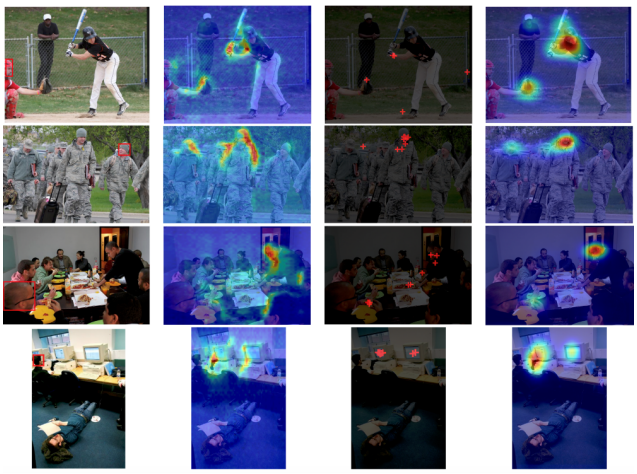

Figure 9. From left to right, we show RGB with given person highlighted, A_i, human annotation and G_i. We can see that human annotations and our predictions share multi-modal property.

Method	Supervision			Evaluation Metric		
	[6]	[18]	Video	AUC ↑	Dist. ↓	AP↑
Random [6]	✓			.505	.458	.621
Center [6]	✓			.728	.326	.624
Chong [5]	✓			.830	.193	.705
Video* [6]	✓			.854	.147	.848
Video [6]	✓		✓	.860	.134	.853
Fang [9]	✓	✓		**.905**	**.108**	**.896**
Ours*	✓			.872	.167	-
Ours	✓			.885	.120	.869

Table 3. Evaluation on the VideoAttentionTarget dataset [6]. Baseline performances are from [6, 9].

dataset to train. It is also interesting to see that even without training on VideoAttentionTarget, Ours* generalizes well and achieves satisfactory results.

5. Conclusion

We propose a novel method for gaze target detection. Our method explicitly models 3D geometry from single RGB image by reconstructing a given scene with 3D point clouds and effectively leverages such information later in ESCNet. To achieve that, we introduce an intermediate representation, or a probability map of front-most 3D points being viewed, and incorporate 3D gaze and scene contextual cues to further regulate the final gaze position. We show that such representation not only offers meaningful understanding of 3D geometry but also allows deep supervision. Our results on two datasets showcase our advantages over existing methods and even human performance.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (62125201, 62020106007, 61836002).

References

[1] Jun Bao, Buyu Liu, and Jun Yu. The story in your eyes: An individual-difference-aware model for cross-person gaze estimation. *arXiv preprint arXiv:2106.14183*, 2021. 2

[2] Jixu Chen and Qiang Ji. 3d gaze estimation with a single camera without ir illumination. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008. 2

[3] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020. 2

[4] 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. 2

[5] E. Chong, Nataniel Ruiz, Y. Wang, Y. Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *ECCV*, 2018. 7, 8

[6] 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. 1, 2, 4, 6, 7, 8

[7] Stefania Cristina and Kenneth P Camilleri. Model-based head pose-free gaze estimation for assistive communication. *Computer Vision and Image Understanding*, 149:157–170, 2016. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[9] 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. 1, 2, 4, 6, 7, 8

[10] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012. 1

[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. 2

[12] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 7

[13] R Hartley and A Zisserman. Multiple view geometry in computer. *Vision, 2nd ed., New York: Cambridge*, 2003. 2, 5

[14] 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 (CVPR)*, June 2016. 7

[15] Takahiro Ishikawa. Passive driver gaze tracking with active appearance models. 2004. 2

[16] Ziyu Jiang, Buyu Liu, Samuel Schulter, Zhangyang Wang, and Manmohan Chandraker. Peek-a-boo: Occlusion reasoning in indoor scenes with plane representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–121, 2020. 2

[17] 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. 6

[18] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, , and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 7, 8

[19] 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. 2

[20] Chen-Yu Lee, Vijay Badrinarayanan, Tomasz Malisiewicz, and Andrew Rabinovich. Roomnet: End-to-end room layout estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4865–4874, 2017. 2

[21] 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. 2, 7

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[23] Buyu Liu, Bingbing Zhuang, Samuel Schulter, Pan Ji, and Manmohan Chandraker. Understanding road layout from videos as a whole. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4414–4423, 2020. 2

[24] Miaomiao Liu, Xuming He, and Mathieu Salzmann. Building scene models by completing and hallucinating depth and semantics. In *European Conference on Computer Vision*, pages 258–274. Springer, 2016. 2

[25] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Inferring human gaze from appearance via adaptive linear regression. In *2011 International Conference on Computer Vision*, pages 153–160. IEEE, 2011. 2

[26] 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

[27] Benoît Massé, Silèye Ba, and Radu 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, 2017. 1

[28] L. Murthy and P. Biswas. Appearance-based gaze estimation using attention and difference mechanism. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3137–3146, 2021. 2

[29] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 721–738, 2018. 2

[30] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 7

[31] Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *NIPS*, 2015. 1, 2, 3, 4, 6, 7, 8

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 7

[33] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014. 2

[34] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 2

[35] Erroll Wood and Andreas Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 207–210, 2014. 2

[36] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[37] Fengting Yang and Zihan Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[38] Yi Yang, Sam Hallman, Deva Ramanan, and Charless Fowlkes. Layered object detection for multi-class segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3113–3120. IEEE, 2010. 2

[39] Yu Yu, Gang Liu, and Jean-Marc Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[40] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12632–12641, 2021. 2

[41] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2018. 2

[42] 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. 2

[43] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 2

[44] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *International Journal of Computer Vision*, 128(5):1076–1100, 2020. 2

[45] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *European Conference on Computer Vision*, pages 550–566. Springer, 2020. 7

[46] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020. 3, 7