Weakly-Supervised Physically Unconstrained Gaze Estimation

Rakshit Kothari[1,2*] Shalini De Mello[1] Umar Iqbal[1]
Wonmin Byeon[1] Seonwook Park[3] Jan Kautz[1]
[1]NVIDIA [2]Rochester Institute of Technology [3]Lunit Inc.
rsk3900@rit.edu; spark@lunit.io
{shalinig, uiqbal, wbyeon, jkautz}@nvidia.com

Abstract

A major challenge for physically unconstrained gaze estimation is acquiring training data with 3D gaze annotations for in-the-wild and outdoor scenarios. In contrast, videos of human interactions in unconstrained environments are abundantly available and can be much more easily annotated with frame-level activity labels. In this work, we tackle the previously unexplored problem of weakly-supervised gaze estimation from videos of human interactions. We leverage the insight that strong gaze-related geometric constraints exist when people perform the activity of "looking at each other" (LAEO). To acquire viable 3D gaze supervision from LAEO labels, we propose a training algorithm along with several novel loss functions especially designed for the task. With weak supervision from two large scale CMU-Panoptic and AVA-LAEO activity datasets, we show significant improvements in (a) the accuracy of semi-supervised gaze estimation and (b) cross-domain generalization on the state-of-the-art physically unconstrained in-the-wild Gaze360 gaze estimation benchmark. We open source our code at https://github.com/NVlabs/weakly-supervised-gaze.

1. Introduction

Much progress has been made recently in the task of remote 3D gaze estimation from monocular images, but most of these methods are constrained to largely frontal subjects viewed by cameras located within a meter of them [46, 20]. To go beyond frontal faces, a few recent works explore the more challenging problem of so-called "physically unconstrained gaze estimation", where larger camera-to-subject distances and higher variations in head pose and eye gaze angles are present [17, 44, 8]. A significant challenge there is in acquiring training data with 3D gaze labels, generally and more so outdoors. Fortunately, several 3D gaze datasets with large camera-to-subject distances and variability in head pose have been collected recently in indoor laboratory environments using specialized multi-cameras setups [43, 8, 44, 28]. In contrast, the recent Gaze360 dataset [17] was collected both indoors and outdoors, at greater distances to subjects. While the approach of Gaze360 advances the field significantly, it nevertheless requires expensive hardware and many co-operative subjects and hence can be difficult to scale.

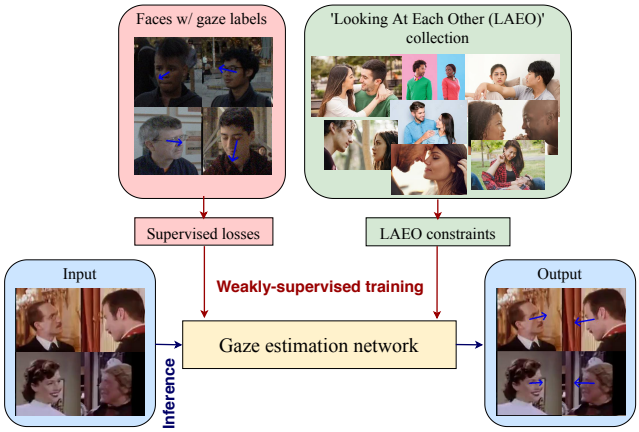

Figure 1. Overview of our weakly-supervised gaze estimation approach. We employ large collections of videos of people "looking at each other" (LAEO) curated from the Internet without any explicit 3D gaze labels, either by themselves or in a semi-supervised manner to learn 3D gaze in physically unconstrained settings.

Recently "weakly-supervised" approaches have been demonstrated on various human perception tasks, such as body pose estimation via multi-view constraints [35, 14], hand pose estimation via bio-mechanical constraints [37], and face reconstruction via differentiable rendering [6]. Nevertheless, little attention has been paid to exploring methods with weak supervision for frontal face gaze estimation [42] and none at all for physically unconstrained gaze estimation. Eye gaze is a natural and strong non-verbal form of human communication [27]. For instance, babies detect and follow a caregiver's gaze from as early as four months of age [38]. Consequently, videos of hu-

*Rakshit Kothari was an intern at NVIDIA during the project.

man interactions involving eye gaze are commonplace and are abundantly available on the Internet [10]. Thus we pose the question: *"Can machines learn to estimate 3D gaze by observing videos of humans interacting with each other?"*.

In this work, we tackle the previously unexplored problem of weakly supervising 3D gaze learning from videos of human interactions curated from the Internet (Fig. 1). We target the most challenging problem within this domain of physically unconstrained gaze estimation. Specifically, to learn 3D gaze we leverage the insight that strong gaze-related geometric constraints exist when people perform the commonplace interaction of "looking at each other" (LAEO), *i.e.*, the 3D gaze vectors of the two people interacting are oriented in opposite directions to each other. Videos of the LAEO activity can be easily curated from the Internet and annotated with frame-level labels for the presence of the LAEO activity and with 2D locations of the persons performing it [26, 25]. However, estimating 3D gaze from just 2D LAEO annotations is challenging and ill-posed because of the depth ambiguity of the subjects in the scene. Furthermore, naively enforcing the geometric constraint of opposing gaze vector predictions for the two subjects performing LAEO is, by itself, insufficient supervision to avoid degenerate solutions while learning 3D gaze.

To solve these challenges and to extract viable 3D gaze supervision from weak LAEO labels, we propose a training algorithm that is especially designed for the task. We enforce several scene-level geometric 3D and 2D LAEO constraints between pairs of faces, which significantly aid in accurately learning 3D gaze information. While training, we also employ a self-training procedure and compute stronger pseudo 3D gaze labels from weak noisy estimates for pairs of faces in LAEO in an uncertainty-aware manner. Lastly, we employ an aleatoric gaze uncertainty loss and a symmetry loss to supervise learning. Our algorithm operates both in a purely weakly-supervised manner with LAEO data only or in a semi-supervised manner along with limited 3D gaze-labeled data.

We evaluate the real-world efficacy of our approach on the large physically unconstrained Gaze360 [17] benchmark. We conduct various within- and cross-dataset experiments and obtain LAEO labels from two large-scale datasets: (a) the CMU Panoptic [16] with known 3D scene geometry and (b) the in-the-wild AVA-LAEO activity dataset [25] containing Internet videos. We show that our proposed approach can successfully learn 3D gaze information from weak LAEO labels. Furthermore, when combined with limited (in terms of the variability of subjects, head poses or environmental conditions) 3D gaze-labeled data in a semi-supervised setting, our approach can significantly help to improve accuracy and cross-domain generalization. Hence, our approach not only reduces the burden of acquiring data and labels for the task of physically unconstrained gaze estimation, but also helps to generalize better for diverse/naturalistic environments.

To summarize, our key contributions are:

- We propose a novel weakly-supervised framework for learning 3D gaze from in-the-wild videos of people performing the activity of "looking at each other". To our understanding, we are the first to employ videos of humans interacting to supervise 3D gaze learning.
- To effectively derive 3D gaze supervision from weak LAEO labels, we introduce several novel training objectives. We learn to predict aleatoric uncertainty, use it to derive strong pseudo-3D gaze labels, and further propose geometric LAEO 3D and 2D constraints to learn gaze from LAEO labels.
- Our experiments on the Gaze360 benchmark show that LAEO data can effectively augment data with strong 3D gaze labels both within and across datasets.

2. Related Work

3D Gaze Estimation Recent developments in remote gaze estimation increasingly benefit from large-scale datasets with gaze direction [46, 9, 36, 8] or target [20, 13] labels. While earlier methods study the effect of different input facial regions [20, 47, 8, 45], later methods attempt to introduce domain-specific insights into their solutions. For example by encoding the eye-shape into the learning procedure [30, 31, 42, 39], or by considering the dependency between head orientation and gaze direction [48, 32, 40], or modelling uncertainty or random effects [41, 2, 17]. Other works propose few-shot adaptation approaches for improving performance for end-users [29, 22, 12, 1, 23]. However, most such approaches restrict their evaluations to screen-based settings (with mostly frontal faces and subjects located within 1m of the camera) due to limitations in the diversity of available training datasets.

Recently proposed datasets such as RT-GENE [8], HUMBI [43], and ETH-XGaze [44] attempt to allow for gaze estimation in more physically unconstrained settings such as from profile faces of subjects located further from the camera. As complex multi-view imaging setups are required, these datasets are inevitably collected in controlled laboratory conditions. A notable exception is Gaze360 [17], which uses a panoramic camera for collecting data from multiple participants at once, both outdoors and indoors. Yet, such collection methods are still difficult to scale compared to data sourced from the web, or via crowd-sourced participation such as done for the GazeCapture dataset [20].

In terms of learning a generalized gaze estimator using only small amounts of labeled data (without supervised pre-training), Yu *et al*. [42] are the only prior art. However, their method is restricted to mostly frontal faces and assumes little to no movement of the head between pairs of samples

from a given participant – an assumption that does not hold in less constrained settings.

Gaze Following and Social Interaction Labels Given an image with a human, gaze following concerns the prediction of the human's gaze target position. Performing this task with deep neural networks was initially explored by Recasens *et al*. [33], with extensions to time sequence data and multiple camera views in [34]. Chong *et al*. [3] improve performance on the static gaze following task further by jointly training to predict 3D gaze direction using the EYEDIAP dataset [9], and by explicitly predicting whether the target is in frame. This work is also extended to video data in [4]. Much like the task of physically unconstrained gaze estimation, gaze following also involves viewing human subjects in all head poses, in diverse environments, and from larger distances. However, gaze following datasets are complex to annotate, and do not lend themselves well to the task of learning to predict 3D gaze due to the lack of scene and object geometry information.

Alternatively, weak annotations for gaze-based interaction exist in the form of social interaction labels. One such condition is the commonplace "looking at each other" condition, also known as LAEO [26], where a binary label is assigned to pairs of human heads for when they are gazing at each other. This is a simpler quantity to annotate compared to mutual attention or visual focus of attention. The recently published AVA-LAEO dataset [25] is an extension of the AVA dataset [10] and demonstrates the ease of acquiring such annotations for existing videos. To the best of our knowledge, we are the first to show that social interaction labels such as LAEO can be used for weakly-supervised gaze estimation. Furthermore, adding LAEO-based constraints and objectives consistently improves performance in cross-dataset and semi-supervised gaze estimation, further validating the real-world efficacy of our approach.

3. Weakly-supervised Gaze Learning

3.1. Problem Definition and Motivation

Our goal is to supervise 3D gaze learning with weak supervision from in-the-wild videos of humans "looking at each other". Such scenes contain the LAEO constraint, *i.e.*, the 3D gazes of the two subjects are oriented along the same line, but in opposite directions to each other. We specifically target the challenging task of physically unconstrained gaze estimation where large subject-to-camera distances, and variations in head poses and environments are present. We assume that we have a large collection of videos containing LAEO activities available to us which can be acquired, for example, by searching the web with appropriate textual queries. We further assume that, by whatever means, the specific frames of a longer video sequence containing the LAEO activity have been located and that the 2D bounding boxes of the pair of faces in the LAEO condition are also available. We refer to these labels collectively as the "LAEO labels".

Acquiring LAEO data is a relatively quick and cost effective way to curate lots of diverse training data. Nevertheless, Internet videos with LAEO labels cannot provide precise 3D gaze supervision. This is because, for such videos neither the scene's precise geometry, nor the camera's intrinsic parameters are known *a priori*. Moreover, trivially enforcing the simple LAEO constraint of requiring the predicted gaze estimates of the two individuals to be opposite to each other is not sufficient for learning gaze. It quickly leads to degenerate solutions.

To address these various challenges, we design a novel weakly-supervised learning framework for 3D gaze estimation from LAEO data. Specifically, we propose a number of novel geometric scene-level LAEO losses, including a 3D and a 2D one, that are applied to pairs of faces in LAEO. For individual face inputs we also use an aleatoric gaze loss [18], which computes gaze uncertainty, along with a self-supervised symmetry loss. We further propose an uncertainty-aware self-training procedure to generate 3D gaze pseudo ground truth labels from pairs of faces exhibiting LAEO. Our training framework operates in two configurations: (a) a purely weakly-supervised one with LAEO data only and (b) a semi-supervised one, where LAEO data is combined with 3D gaze-labeled data.

3.2. Solution Overview

Our overall framework for weakly-supervised 3D gaze learning from LAEO data is shown in Fig. 2. We wish to train the function $\mathcal{F}(\mathbf{I}, \theta)$ with weights θ to estimate gaze by providing video sequences of pairs of people exhibiting LAEO. Inspired by [17], our gaze estimation network $\mathcal{F}(\mathbf{I}, \theta)$ consists of a ResNet-18 backbone followed by two bi-directional LSTM layers and a fully-connected (FC) layer, which estimates a gaze value $\hat{g} = \{\hat{g}_\theta, \hat{g}_\phi\}$ along with an uncertainty value $\hat{\sigma}$ corresponding to the central image in a sequence of 7 consecutive input frames. Here $\hat{g}_\theta$ and $\hat{g}_\phi$ indicate the estimates for the gaze pitch and yaw angles, respectively. In addition to this temporal version of our network, we also explore a *static* variant, which takes a single image as input and bypasses the LSTM layers to directly connect the output of the backbone CNN to the FC layer.

For LAEO data, the input to our network is a pair of head crops of size $224 \times 224 \times 3$ each containing one of the two faces that exhibit LAEO along with the original scene image. No 3D gaze labels are available during training with LAEO data. If data containing explicit 3D gaze labels is additionally available for semi-supervised training, we extract single head crops from the scene images and input them along with their known ground truth 3D gaze labels into the network for training.

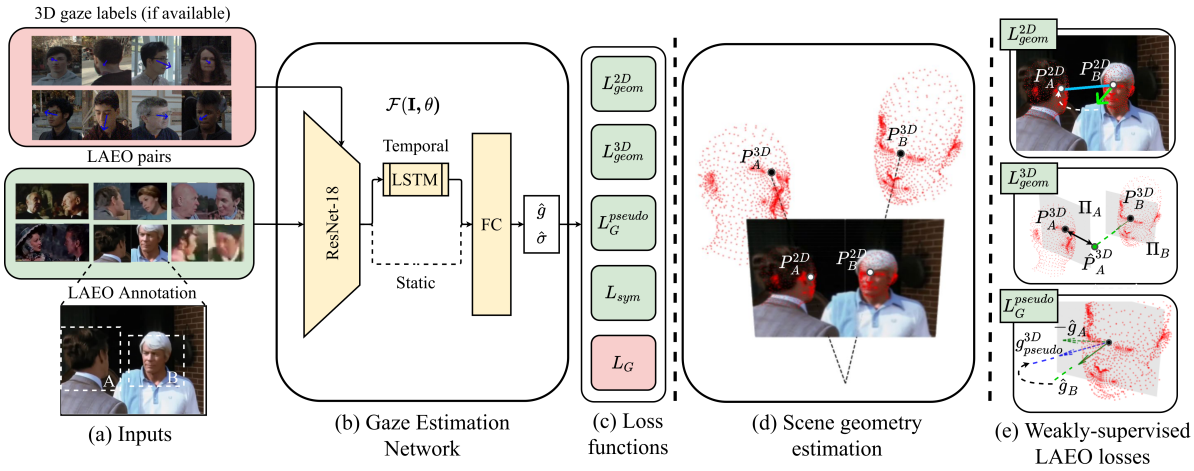

Figure 2. An overview of our weakly-supervised approach to learning 3D gaze from the "looking at each other" human activity videos. From left to right, we show (a) the inputs to our gaze estimation network, *i.e.*, pairs of head crops in LAEO with their scene images for weakly-supervised training and optionally single head crops with gaze labels for semi-supervised training (if available); (b) our gaze estimation network, which predicts gaze $\hat{g}$ and its uncertainty $\hat{\sigma}$; (c) the various weakly-supervised and fully-supervised losses used for training; (d) estimation of scene geometry for in-the-wild LAEO videos acquired from the web including that of the 2D and 3D positions of the cyclopean eyes of the subject pairs in LAEO used to compute the LAEO losses; and (e) details of our proposed LAEO losses including the geometric 2D LAEO loss (L^{2D}_{geom}), geometric 3D LAEO loss (L^{3D}_{geom}) and the pseudo gaze loss (L^{pseudo}_{G}).

3.3. Loss Functions

We employ several end-to-end differentiable geometric loss functions, which are derived from the LAEO constraint to supervised 3D gaze learning. These include two scene-level geometric 2D and 3D LAEO losses. We start by describing our technique for scene geometry estimation and 3D gaze origin determination for in-the-wild videos and then describe our geometric LAEO losses. We then describe our uncertainty-aware 3D gaze pseudo labeling procedure, followed by two additional losses – the aleatoric gaze and symmetry losses that are applied to individual face inputs.

Scene Geometry Estimation The geometric LAEO loss functions can only be computed in a coordinate system common to both subjects, *i.e.*, the camera coordinate system. For Internet videos, we cannot reliably recover camera parameters or the subjects' 3D poses. So we instead derive approximate values for them. We approximate the camera focal parameter f to be the size of the larger image dimension in pixels. The principal point is assumed to be at the center of the image. We detect the 2D facial landmarks of a subject using AlphaPose [7] and refer to the midpoint of their left and right eye pixel locations as their "2D cyclopean eye" $P^{2D} = (x, y)$. We assume it to be the point from where gaze originates for a subject on the 2D image plane. To find its 3D counterpart, *i.e.*, the 3D cyclopean eye P^{3D}, we also estimate depth z per subject and back-project P^{2D} to 3D as $(zx/f, zy/f, z)$. This procedure ensures that P^{2D} and P^{3D} lie on the same projection line originating from the camera's center.

To recover depth z of each subject, we first estimate their 2D-3D correspondences using DensePose [11]. We use the predicted 2D facial key-points [7] and an average gender neutral 3D SMPL head model [24] to compute the 3D transformation required to fit the 3D head model to a particular subject using PnP [21]. This allows for an estimation of the 3D head model's location and orientation in the camera coordinate system, which in turn provides us with depth estimates in meters (see Fig. 2) for each subject. Specifically, we utilize the depth z value of the mid points of the left and right eyes of the fitted 3D head model to recover depth of each subject. The end result is a shared 3D coordinate system for both subjects under LAEO (see Fig. 2). In Sec. D.3 of the supplementary we further discuss the effect of our various approximations employed to compute the scene geometry on the reliability of 3D gaze estimates derived from LAEO data.

Geometric 2D LAEO Loss For two subjects A and B in LAEO, the projections of their predicted 3D gaze vectors onto the scene image plane, should lie along the line joining their 2D cyclopean eyes P^{2D}_A and P^{2D}_A (see Fig. 2). This intuition forms the basis of our geometric 2D LAEO loss L^{2D}_{geom}. To compute this loss, we estimate the gaze angles $\hat{g}_A$ for subject A in LAEO by forward propagating their head crop image I_A through $\mathcal{F}(\mathbf{I}, \theta)$. We then transform it to a 3D unit gaze vector $\hat{g}^{3D}_A$ originating from subject A's 3D cyclopean eye P^{3D}_A in the camera coordinate system. Next, we forward project $\hat{g}^{3D}_A$ onto the observed scene image as the 2D gaze vector $\hat{g}^{2D}_A$ (see Fig. 2). To compute L^{2D}_{geom}, we compute the angular cosine distance between two 2D unit vectors in the image plane: one along $\hat{g}^{2D}_A$ and

another one along the line joining P_A^{2D} and P_B^{2D}. We repeat this process for subject B and average both losses to obtain the final loss L_{geom}^{2D}.

Note, however, that L_{geom}^{2D} on its own cannot fully resolve the depth ambiguity present in videos obtained from the Internet and hence is not sufficient to learn 3D gaze (see Table 1), but when combined with the other LAEO losses it helps to improve overall gaze estimation accuracy (see Sec. B.4 in supplementary). Thus, we additionally propose a geometric 3D LAEO loss which helps to resolve depth ambiguities and aids in learning 3D gaze more accurately. We describe it next.

Geometric 3D LAEO Loss The geometric 3D LAEO loss, L_{geom}^{3D}, explicitly provides 3D directional information to supervise gaze learning. We formulate it to enforce that the estimated 3D gaze vector originating from the cyclopean eye P_B^{3D} of subject B in LAEO, must intersect the viewed subject A's 3D cyclopean eye P_A^{3D} (see Fig. 2). To achieve this, we first estimate the 3D facial plane Π_A of the viewed subject A, and place it at their 3D cyclopean eye location P_A^{3D} perpendicular to their heading vector. We define the heading vector as the line joining the 3D midpoint of a subject's outer most 3D ear points, and 3D nose tip obtained from the fitted SMPL head model. Then the geometric 3D LAEO constraint for subject B is given by $||\hat{P}_A^{3D}$ - $P_A^{3D}||$, where P_A^{3D} is subject A's 3D cyclopean eye position and $\hat{P}_A^{3D}$ is the intersection of subject B's 3D gaze vector $\hat{g}_B^{3D}$ with subject A's face plane Π_A (see Fig. 2). Here $||.||$ denotes Euclidean distance. We repeat this process for subject A and average the losses computed for both subjects to obtain the final loss L_{geom}^{3D}. Empirically we find that our formulation for L_{geom}^{3D} performs better than an alternate cosine angle-based version (see Sec. B.3 in supplementary).

Pseudo Gaze LAEO Loss The LAEO activity also provides us with the self-supervised constraint that the ground truth 3D gaze vectors of two individuals A and B in LAEO, are oriented along the same 3D line, but in opposite directions to each other, *i.e.*, $g_A^{3D} = -g_B^{3D}$. Hence, we leverage it in a self-training procedure and compute gaze pseudo ground truth labels for a pair of LAEO subjects continually while training. We observe that the LAEO activity often results in a clear frontal view of one subject while the other subject is turned away (see examples in Fig. 1 and Fig. 2). Moreover, gaze estimation errors generally increase with extreme head poses where features such as the eyes are less visible (see Fig. 2 in the supplementary for a plot of gaze error versus gaze yaw). For example, in the extreme case of looking from behind a subject, facial features become completely occluded.

We find that the uncertainty measure estimated by our network is well correlated with gaze error (with a Spearman's rank correlation coefficient of value of 0.46). So to derive the gaze pseudo ground truth for a pair of faces in LAEO, we use the uncertainty measure to weigh more heavily the more reliable (less uncertain) of the two gaze estimates for a LAEO pair. Specifically, let $\{\hat{g}_A^{3D}, \hat{\sigma}_A\}$ and $\{\hat{g}_B^{3D}, \hat{\sigma}_B\}$ be the predicted 3D gaze vectors and their angular uncertainty values (in a common 3D coordinate system) for a pair of input face crops in LAEO, $\mathbf{I_A}$ and $\mathbf{I_B}$, respectively. We compute the pseudo 3D gaze ground truth label g_{pseudo}^{3D} for faces A and B as a weighted combination of their estimated 3D gaze vectors as:

$$g_{pseudo}^{3D} = \mathrm{w_A}\hat{g}_A^{3D} + \mathrm{w_B}(-\hat{g}_B^{3D}), \tag{1}$$

where we compute $\mathrm{w_A}$ and $\mathrm{w_B}$ from the angular uncertainty values $\hat{\sigma}_A$ and $\hat{\sigma}_B$ as $\mathrm{w_A} = {\hat{\sigma}_B}/{\hat{\sigma}_A+\hat{\sigma}_B}$ and $\mathrm{w_B} = {\hat{\sigma}_A}/{\hat{\sigma}_A+\hat{\sigma}_B}$ predicted by the gaze network. We further compute cosine distances between each LAEO subjects' predicted gaze vectors $\hat{g}^{3D}$ and their respective pseudo ground truth values g_{pseudo}^{3D} and $-g_{pseudo}^{3D}$. We average the cosine distances computed for both subjects to obtain the final L_G^{pseudo} loss. We find that this formulation of L_G^{pseudo} is superior to other variants of it (see Sec. B.2 in supplementary).

Aleatoric Gaze Loss We use an aleatoric loss function L_G to supervise gaze estimation of individual face inputs, which regresses both the predicted gaze value and its uncertainty. This gaze uncertainty is helpful in deriving pseudo ground truths for pairs of faces in LAEO as described in the previous section. Aleatoric uncertainty models the distribution of the estimated gaze angles as a parametric Laplacian function and hence our gaze network $\mathcal{F}(\mathbf{I}, \theta)$ predicts their estimated mean $\{\hat{g}_\theta, \hat{g}_\phi\}$ and absolute deviation $\hat{\sigma}$ values. We supervise the network by minimizing the negative log-likelihood of observing the ground truth gaze value $\{g_\theta, g_\phi\}$ w.r.t. to this predicted Laplacian distribution as:

$$\begin{aligned} L_G^\theta &= \log(\hat{\sigma}) + \frac{1}{\hat{\sigma}}|\hat{g}_\theta - g_\theta| \\ L_G^\phi &= \log(\hat{\sigma}) + \frac{1}{\hat{\sigma}}|\hat{g}_\phi - g_\phi| \\ L_G &= L^\phi + L^\theta. \end{aligned} \tag{2}$$

In practice, we predict the logarithm of the absolute deviation $\log(\hat{\sigma})$ from our network. This formulation has been shown to be numerically stable and avoids a potential division by zero [18]. Note that previously, in [17], the authors similarly employed a pinball loss to estimate the uncertainty of gaze predictions. We find that, in comparison to the pinball loss, the aleatoric loss improves the baseline accuracy of gaze estimation (see Sec. B.1 in supplementary).

Symmetry Loss We also exploit the left-right symmetry inherent to the gaze estimation task to enforce another self-supervised gaze symmetry loss L_{sym}. Specifically, we estimate gaze angles for an input face image $\mathbf{I}$ as $\hat{g} = \{\hat{g}_\theta, \hat{g}_\phi\}$, reverse the sign of its predicted gaze yaw angle to produce

the altered prediction $\hat{g}^* = \{\hat{g}_\theta, -\hat{g}_\phi\}$ and use this altered gaze estimate as the ground truth to supervise the predicted gaze, using the aleatoric loss, for a horizontally flipped (mirrored) version $\mathbf{I}^*$ of the input face image as:

$$L_{sym} = L_G(\mathcal{F}(\mathbf{I}^*, \theta), \hat{g}^*). \tag{3}$$

We repeat this process for the horizontally flipped image and average the two resultant losses. Note that here the gaze angles are assumed to be in a normalized eye coordinate system as described in [17], whose z axis passes through each subject's 3D cyclopean eye position P^{3D}. This loss prevents network over-fitting while improving accuracy of gaze estimation (see Sec. B.1 in supplementary).

3.4. Training

We adopt two training paradigms: purely weakly-supervised training with LAEO data only or semi-supervised training where LAEO data augments data containing explicit 3D gaze labels. In both conditions, we initialize the ResNet-18 backbone of our model with weights pre-trained using ImageNet [5]. We initialize the LSTM module and FC weights using a normal distribution. For semi-supervised training, we first train our model to convergence with images containing explicit 3D gaze labels only and then add weakly-supervised images with LAEO labels and continue training jointly to convergence. We fix the parameters of the batch normalization layers during initialization to those found in the ImageNet pre-trained weights. We optimize the model using the following objective function:

$$\begin{aligned} L &= L_G + \alpha L_{sym} + \beta L_{LAEO}, \\ L_{LAEO} &= (L^{3D}_{geom} + L^{2D}_{geom} + L^{pseudo}_{G}). \end{aligned} \tag{4}$$

Here, α and β are scalar weights, which slowly ramp up the contribution of the symmetry and LAEO losses, respectively. The ramp operation is formulated as $(\lceil i/T \rceil^1)$ where i is the smallest iterative step to update our model while T is a threshold. We set T_α as 3000 and T_β as 2400. In experiments, which do not involve any gaze supervision, β is always fixed at 1 and L_G is not included. We use a batch-size of 80 frames/sequences to train our static/temporal gaze estimation network. We use a fixed learning rate of 10^{-4} with the ADAM optimizer [19].

4. Experiments

Here we evaluate the real-world performance of our method in the fully weakly-supervised or semi-supervised settings for the task of physically unconstrained gaze estimation [17]. We perform extensive experiments within and across datasets. Besides gaze estimation, in Sec. A of the supplementary, we also show the utility of adding LAEO labels to the task of in-the-wild visual target attention prediction [4] in a semi-supervised setting.

LAEO Datasets We employ two LAEO datasets – CMU Panoptic [16] and AVA [10, 25]. CMU Panoptic [15] is collected with a multiple-camera system installed in a large indoor dome, wherein subjects perform various activities. It does not contain LAEO annotations but contains the subjects' 3D body-joint locations and camera calibration information, which we directly use in our experiments. From video sequences containing the *haggling* activity, we extract clips with the LAEO activity present via a semi-automatic procedure (described Sec. D.1 of supplementary). This results in over 800k pairs of faces extracted from 485 unique subjects. For our experiments, we only utilize images from cameras that are parallel to the ground plane.

To acquire LAEO data from in-the-wild Internet videos, we leverage the large scale AVA human activity dataset [10] with LAEO annotations [25] provided by Marin-Jimenez *et al*. (called "AVA-LAEO"). It consists of annotated head bounding box pairs under LAEO in select frames across multiple video sequences, resulting in a wide variety of faces, backgrounds and lighting conditions. Unlike CMU Panoptic, AVA-LAEO does not provide access to accurate camera parameters or 3D human poses. We estimate the subjects' 3D poses using DensePose [11] and AlphaPose [7] (described in Sec. 3.3 and Sec. D.2 in supplementary). In all, this dataset contains 13,787 sequences of pairs of faces in LAEO.

Gaze Datasets We validate the efficacy of our weakly-supervised approach on the large-scale physically unconstrained in-the-wild Gaze360 [17] dataset. It contains explicit 3D gaze labels and large variations in subject head poses and gaze angles, and lighting conditions and backgrounds. Its images are acquired in both indoor and outdoor environments using a Ladybug multi-camera system. It contain 127K training sequences from 365 subjects. For semi-supervised training, we additionally use two large-scale gaze datasets with known 3D gaze ground-truth – GazeCapture [20] and ETH-XGaze [44]. GazeCapture contains nearly 2M frontal face images of 1474 subjects acquired in unconstrained environmental conditions. ETH-XGaze, on the other hand, was acquired indoors with controlled lighting on a standard green background with a multi-view camera system. It contains 756K frames of 80 subjects.

The gaze distribution plots of all these datasets and their example face images are shown in Fig. 3. For GazeCapture and ETH-XGaze, we use the normalization procedure described in [47] to create normalized face crops. For all other datasets, we employ the procedure described in [17] to create normalized head crops. For all evaluations, we report the angular error (in degrees) between the estimated and ground truth unit gaze vectors, averaged across the corresponding test dataset.

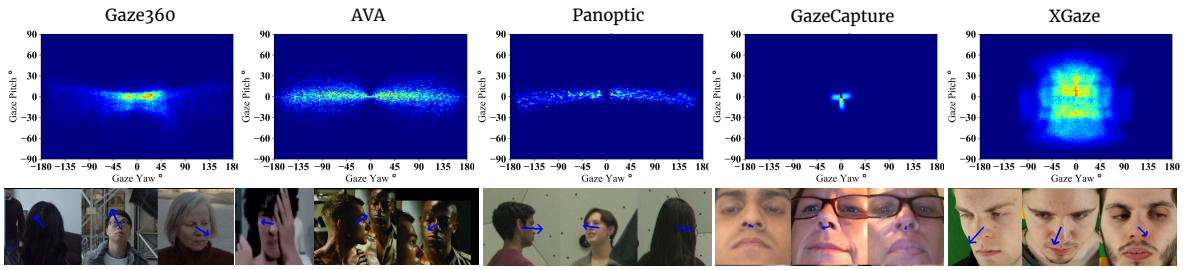

Figure 3. **Top** Gaze direction distribution of the Gaze360 [17], AVA-LAEO [25, 10], CMU Panoptic [15], GazeCapture [20] and ETH-XGaze [44] datasets. Note that here the approximate gaze for CMU Panoptic and AVA-LAEO is computed by joining the LAEO pair of subjects' 3D cyclopean eye locations. **Bottom** Example face or head crops (if available) from the individual datasets.

4.1. Ablation Study

To verify the contributions of our individual losses, we conduct a purely weakly-supervised cross-dataset ablation study. We train our method with the CMU Panoptic or AVA-LAEO datasets and evaluate performance on the Gaze360 dataset's test partition. Table 1 highlights the effect of the various weakly-supervised LAEO losses in this cross-dataset setting. All values reported are for the case when the symmetry loss was used by default. We train two configurations of our gaze estimation model – (a) a temporal version, which accepts 7 frames as input, and (b) a static variant, which predicts gaze from a single input frame.

We observe that among the individual weakly-supervised losses, L_G^{pseudo} and L_{geom}^{2D} on their own or together, result in degenerate solutions. This is not surprising as it highlights the effects of depth ambiguity (see Sec. 3.3). Strong supervision can be provided by explicitly constraining the estimated gaze to intersect a 3D target, in our case, the viewed subject's head in the LAEO condition. This can be seen from the fact that L_{geom}^{3D} significantly improves over its degenerate counterparts. We observe that the best performance is achieved by utilizing a combination of L_{geom}^{3D}, L_{geom}^{2D} and the L_G^{pseudo} losses, especially with the real-world AVA-LAEO dataset where the scene geometry is not known. We also find that removing the symmetry loss increases the overall gaze error of our best (temporal) model to 27.9° from 25.9° for CMU Panoptic and to 27.9° from 26.3° for AVA-LAEO (not listed in Table 1). We provide additional ablation studies to explore the effect of the aleatoric and symmetric losses; other variants of the L_G^{pseudo} and L_{geom}^{3D} losses; and the utility of L_{geom}^{2D} in Sec. B of the supplementary.

	Temporal		Static	
Loss functions	Pano.	AVA	Pano.	AVA
L_G^{pseudo}	55.4	52.9	61.9	48.0
L_{geom}^{2D}	58.7	52.4	49.0	46.9
L_{geom}^{3D}	28.0	30.1	31.4	30.4
$L_{geom}^{2D} + L_G^{pseudo}$	55.0	54.1	54.0	51.3
$L_{geom}^{3D} + L_G^{pseudo}$	26.1	27.3	**29.0**	30.6
$L_{geom}^{3D} + L_{geom}^{2D}$	26.9	26.4	31.3	30.8
$L_G^{pseudo} + L_{geom}^{3D} + L_{geom}^{2D}$	**25.9**	**26.3**	31.3	**28.7**

Table 1. An ablation study to evaluate our individual weakly-supervised LAEO losses. The symmetry loss is always used. All numbers reported are using predictions from the temporal and static variants of our gaze estimation model, when evaluated on Gaze360's test set, measured in gaze angular error in degrees. Lower is better.

4.2. Semi-supervised Evaluation

Despite successfully learning to estimate gaze, the performance of our purely weakly-supervised model (trained on the AVA-LAEO dataset and tested on the Gaze360 dataset) lags behind the fully-supervised model on Gaze360's training set [17] as shown in Table 2 (26.3° vs 13.2° for the temporal model). One reason for this discrepancy is the presence of noise in the gaze labels derived from LAEO data (as discussed in Sec. D.3 of the supplementary) and the other is the the existence of domain gap between the AVA-LAEO and Gaze360 datasets. The latter is evident from the gaze distribution plots shown in Fig. 3. LAEO data tends to be biased towards viewing individuals from larger profile angles (see Fig. 1 and Fig. 2) and contains less frontal face data. It also contains less diversity in the head's pitch (up/down rotation).

Hence, in this experiment, we explore a semi-supervised setting, where we evaluate if weakly-supervised LAEO data can successfully augment limited gaze-labeled data and improve its generalization for the task of physically unconstrained gaze estimation in-the-wild. We conduct both cross-dataset and within-dataset experiments. For the cross-dataset experiment, we train our model with several existing datasets other than Gaze360 and test on Gaze360's test partition. For the within-dataset experiment, we train on various subsets of Gaze360's training partition along with LAEO data and evaluate performance on Gaze360's test set. Unlike [44], which evaluates performance on only frontal faces from Gaze360, we evaluate performance on both (a) frontal and (b) all faces from Gaze360's test set (including large profile faces).

Cross-dataset In Table 2, we compare the generalization performance of the GazeCapture and the ETH-XGaze

Within dataset, Gaze + LAEO Labels		
Training Data	Frontal face crops	All head crops
Gaze360 [17]	11.1	13.5
Gaze360	**10.1**	**13.2**
Gaze360 + AVA	10.2	**13.2**
Cross dataset, Gaze + LAEO Labels		
Training Data	Frontal face crops	All head crops
GazeCapture [44]	30.2	-
GazeCapture	29.2	58.2
GazeCapture + AVA	19.5	27.2
ETH-XGaze [44]	27.3	-
ETH-XGaze	20.5	52.6
ETH-XGaze + AVA	**16.9**	**25.0**
Cross dataset, LAEO Labels		
Training Data	Frontal face crops	All head crops
AVA	29.0	26.3
CMU Panoptic	26.0	25.9
CMU Panoptic + AVA	**22.5**	**24.4**

Table 2. Performance evaluation of our temporal model on Gaze360 dataset's test partition with various different training datasets ranging from those containing full gaze supervision (Gaze360, GazeCapture, ETH-XGaze), weak LAEO supervision only (the AVA-LAEO or CMU Panoptic datasets), or their combinations. All reported values are gaze angular errors in degrees (lower is better) on either (a) frontal face crops only or (b) all head crops from Gaze360's test set. Note that the addition of AVA-LAEO to GazeCapture or ETH-XGaze significantly improves their generalization performance on Gaze360.

datasets on Gaze360, with and without weak supervision from AVA-LAEO. Both these supervised gaze datasets, although large, are limited in some respect for the task of physically unconstrained gaze estimation in-the-wild. The GazeCapture dataset contains images acquired indoors and outdoors, but of mostly frontal faces with a narrow distribution of gaze angles (Fig. 3). The ETH-XGaze dataset, on the other hand, has a broad distribution of gaze angles from 80 subjects (Fig. 3), but is captured indoors only.

Table 2 highlights that on including weak gaze supervision from AVA-LAEO, the generalization performances of both GazeCapture and ETH-XGaze on Gaze360, for frontal and all faces is improved. For frontal faces, the addition of AVA-LAEO results in improvements of $7.4°$ for GazeCapture and $3.6°$ for ETH-XGaze. On all head crops, however, this improvement is even more pronounced – $31.0°$ for GazeCapture and $27.6°$ for ETH-XGaze. Fig. 3 shows that the AVA-LAEO dataset complements both the GazeCapture and ETH-XGaze datasets by expanding their underlying distributions via weak gaze labels (see more details in Sec. C of supplementary). In Table 2, we also show the cross-dataset performance of jointly training with CMU Panoptic and AVA-LAEO with their weak gaze labels only. We find that the in-the-wild AVA-LAEO data also slightly

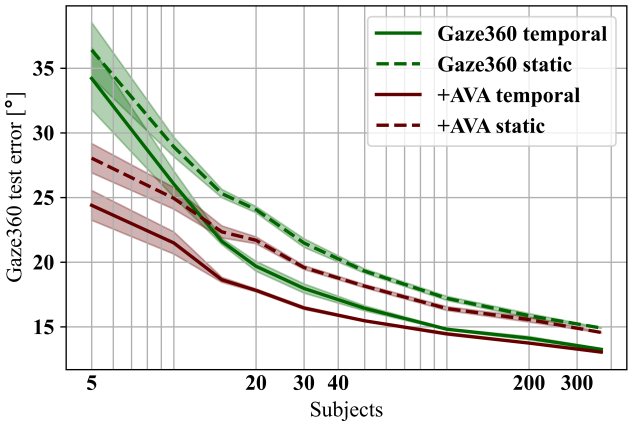

Figure 4. Gaze error (in degrees) on the Gaze360 test set on augmenting a reduced Gaze360 training set (with less subjects) with AVA-LAEO. We vary the number of Gaze360 training subjects along the horizontal axis. The shaded area corresponds to the standard error of the average metric evaluated over 5 repetitions of each experiment performed by picking a different random subset of subjects each time.

improves the generalization performance of the indoor-only CMU Panoptic data on Gaze360. Finally, Table 2 shows that our model also outperforms the previously reported state-of-the-art performances [17, 44] on all benchmarks.

Within-dataset Training data acquired from a larger number of subjects improves generalization of gaze estimators as shown in [20]. However, recruiting more subjects requires additional cost and time. In Fig. 4, we evaluate the performance of training with progressively larger numbers of subjects from Gaze360's training set, without (labeled as "Gaze360" in Fig. 4) and with (labeled as "+AVA" in Fig. 4) AVA-LAEO. We use all available videos of a particular subject during training. We assess both our temporal and static models. For this within-domain semi-supervised setting, we find that training on a small number of subjects from Gaze360 along with weak supervision from AVA-LAEO, offers the same performance as using a larger number of subjects from Gaze360.

5. Conclusion

In this work, we present the first exploration of a weakly-supervised 3D gaze learning paradigm from images/videos of people *looking at each other* (LAEO). This approach is trivially scalable due to the ease of acquiring LAEO annotations from Internet videos. To facilitate the learning of 3D gaze, we propose three training objectives, which exploit the underlying geometry native to the LAEO activity. Through many experiments we demonstrate that our approach is successful in augmenting gaze datasets limited in gaze distributions, subjects, or environmental conditions with unconstrained images of people under LAEO, resulting in improved physically unconstrained gaze estimation in the wild.

References

[1] Zhaokang Chen and Bertram Shi. Offset calibration for appearance-based gaze estimation via gaze decomposition. In *WACV*, pages 270–279, 2020.

[2] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *ECCV*, pages 100–115, 2018.

[3] 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 *ECCV*, pages 383–398, 2018.

[4] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting Attended Visual Targets in Video. In *CVPR*, 2020.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[6] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*, 2019.

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[8] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *ECCV*, pages 334–352, 2018.

[9] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *ACM ETRA*. ACM, Mar. 2014.

[10] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *CVPR*, 2018.

[11] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. In *CVPR*, 2018.

[12] Junfeng He, Khoi Pham, Nachiappan Valliappan, Pingmei Xu, Chase Roberts, Dmitry Lagun, and Vidhya Navalpakkam. On-device few-shot personalization for real-time gaze estimation. In *ICCV Workshops*, Oct 2019.

[13] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: Dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Mach. Vision Appl.*, 28(5-6):445–461, Aug. 2017.

[14] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, pages 5243–5252, 2020.

[15] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, pages 3334–3342, 2015.

[16] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. *TPAMI*, 2019.

[17] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. *ICCV*, pages 6911–6920, 2019.

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *NeurIPS*, pages 5575–5585, 2017.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye Tracking for Everyone. In *CVPR*, June 2016.

[21] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *IJCV*, 81(2):155, 2009.

[22] Erik Lindén, Jonas Sjostrand, and Alexandre Proutiere. Learning to personalize in appearance-based gaze tracking. In *ICCV Workshops*, pages 0–0, 2019.

[23] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *TPAMI*, 2019.

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 34(6):248:1–248:16, Oct. 2015.

[25] Manuel J. Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: Revisiting people looking at each other in videos. *CVPR*, 2019-June(i):3472–3480, 2019.

[26] M. J. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari. Detecting people looking at each other in videos. *IJCV*, 2014.

[27] Allan Mazur, Eugene Rosa, Mark Faupel, Joshua Heller, Russell Leen, and Blake Thurman. Physiological aspects of communication via mutual gaze. *American Journal of Sociology*, 86(1):50–74, 1980.

[28] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *ECCV*, 2020.

[29] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *ICCV*, 2019.

[30] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *ECCV*, pages 721–738, 2018.

[31] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *ACM ETRA*, pages 1–10, 2018.

[32] Rajeev Ranjan, Shalini De Mello, and Jan Kautz. Lightweight head pose invariant gaze tracking. In *CVPR Workshops*, pages 2156–2164, 2018.

[33] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *NeurIPS*, pages 199–207, 2015.

[34] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *ICCV*, pages 1435–1443, 2017.

[35] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *CVPR*, pages 8437–8446, 2018.

[36] B.A. Smith, Q. Yin, S.K. Feiner, and S.K. Nayar. Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction. In *ACM UIST*, pages 271–280, Oct 2013.

[37] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *ECCV*, 2020.

[38] Tricia Striano and Vincent M Reid. Social cognition in the first year. *Trends in cognitive sciences*, 10(10):471–476, 2006.

[39] Kang Wang, Rui Zhao, and Qiang Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In *CVPR*, pages 440–448, 2018.

[40] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *CVPR*, pages 11907–11916, 2019.

[41] Yunyang Xiong, Hyunwoo J Kim, and Vikas Singh. Mixed effects neural networks (menets) with applications to gaze estimation. In *CVPR*, pages 7743–7752, 2019.

[42] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *CVPR*, pages 7314–7324, 2020.

[43] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *CVPR*, pages 2990–3000, 2020.

[44] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *ECCV*, 2020.

[45] Xucong Zhang, Yusuke Sugano, Andreas Bulling, and Otmar Hilliges. Learning-based region selection for end-to-end gaze estimation. In *BMVC*, pages 1–13, 2020.

[46] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, pages 4511–4520, June 2015.

[47] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *CVPR Workshops*, pages 2299–2308, July 2017.

[48] Wangjiang Zhu and Haoping Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *ICCV*, pages 3143–3152, 2017.