# 3DGazeNet: Generalizing 3D Gaze Estimation with Weak-Supervision from Synthetic Views

Evangelos Ververas[1,2], Polydefkis Gkagkos[2], Jiankang Deng[1,2], Michail Christos Doukas[1], Jia Guo[3], and Stefanos Zafeiriou[1]

[1] Imperial College London, UK
[2] Huawei Noah's Ark Lab, UK
[3] InsightFace
https://eververas.github.io/3DGazeNet/

**Abstract.** Developing gaze estimation models that generalize well to unseen domains and in-the-wild conditions remains a challenge with no known best solution. This is mostly due to the difficulty of acquiring ground truth data that cover the distribution of faces, head poses, and environments that exist in the real world. Most recent methods attempt to close the gap between specific source and target domains using domain adaptation. In this work, we propose to train general gaze estimation models which can be directly employed in novel environments without adaptation. To do so, we leverage the observation that head, body, and hand pose estimation benefit from revising them as dense 3D coordinate prediction, and similarly express gaze estimation as regression of dense 3D eye meshes. To close the gap between image domains, we create a large-scale dataset of diverse faces with gaze pseudo-annotations, which we extract based on the 3D geometry of the face, and design a multi-view supervision framework to balance their effect during training. We test our method in the task of gaze generalization, in which we demonstrate improvement of up to 23% compared to state-of-the-art when no ground truth data are available, and up to 10% when they are.

**Keywords:** 3D Gaze Estimation · 3D Eye Mesh · Gaze Generalization

## 1 Introduction

Eye gaze serves as a cue for understanding human behavior and intents, including attention, communication, and mental state. As a result, gaze information has been exploited by a lot of applications of various fields of interest, ranging from medical and psychological analysis [9, 37, 64] to human-computer interaction [4], efficient rendering in VR/AR headset systems [6, 10, 39], virtual character animation [57, 61, 62, 77] and driver state monitoring [34, 50]. When high accuracy is important, data collection under the particular capturing set up is crucial, e.g. specific VR headsets, static screen-camera setups. However, in numerous real-world applications robustness is equally important to high accuracy, e.g. face-unlocking in mobile devices, best frame capturing/selection in group photos and automatic gaze annotation of large datasets for face reenactment.

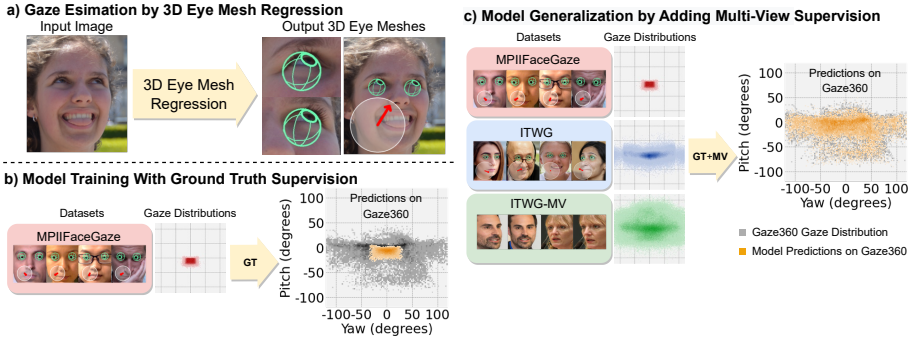



**Fig. 1:** Overview of our method 3DGazeNet. a) We approach 3D gaze estimation as dense 3D eye mesh regression, which is robust against sparse prediction errors. b) Domain generalization is one of the hardest challenges in gaze estimation. Training with common gaze datasets often results in poor cross-dataset performance. c) Our multi-view supervision method employs pseudo-labels from in-the-wild face images to close the gap between controlled and in-the-wild datasets.

Typically, 3D gaze estimation is expressed as a direct mapping between input images and a few pose parameters [12, 42, 52, 70, 82], or sparse representations of the eyes [54, 55, 66]. Nevertheless, it has been shown that unconstrained face and body pose estimation from single images benefits from replacing predicting few pose or shape parameters by directly predicting dense 3D geometry [3, 16, 26, 43, 58]. In this work, we leverage this observation and revise the formulation of gaze estimation as end-to-end dense 3D eye mesh regression, which combined with standard vector regression induces multiple benefits. Existing datasets with ground truth 3D eyes include only images in the IR domain [21], however, IR images cannot be directly employed for RGB-based methods. As 3D eye meshes are not available for most gaze datasets, we define a unified eye representation, i.e. a rigid 3D eyeball template (Fig. 3(a)), which we fit on images based on sparse landmarks and the available gaze labels.

Several gaze datasets have become available in the last decade [20, 22, 35, 42, 52, 59, 60, 79, 81], which have contributed to the recent progress in automatic 3D gaze estimation from monocular RGB images. However, collecting gaze datasets is a costly and challenging process which often restricts them being captured in controlled environments and consisting of limited unique identities, thus lacking variation compared to data from the real world. This causes the most common challenge in gaze estimation, which is cross-domain and in-the-wild generalization. In this work, we propose a method to exploit arbitrary, unlabeled face images to largely increase the diversity of our training data as well as our model's generalization capabilities. To that end, we design a simple pipeline to extract robust 3D gaze pseudo-labels based on the 3D shape of the face and eyes, without having any prior gaze information. Based on recent advancements on weakly-supervised head, body and hand pose estimation [8, 17, 31, 44, 65], we regularize inconsistencies of pseudo-labels, by a geometric constraint which encourages our

model to maintain prediction consistency between multiple synthetic views of the same subject.

Most recent methods attempt to close the gap between diverse image domains using domain adaptation. Commonly, they employ a few samples of the target domain, with [29,53,73] or without [5,7,11,24,27,47,68,70] their gaze labels, to fine-tune an initial model. Although successful, approaches following this scheme require knowledge of the target domain and model re-training, which prohibit their use as plug-n-play methods in real user applications. In contrast, we propose a method to train gaze estimation models that generalize well to unseen and in-the-wild environments without the constraints of domain adaption. Our method can effortlessly be employed by user applications in a plug-n-play fashion.

An overview of our approach, which we name 3DGazeNet, is presented in Fig. 1. We evaluate our method in cross-dataset gaze generalization, showcasing improvements over the state-of-the-art, even by a large margin, and perform ablations over the model components. To summarize, the key contributions of our work are:

- A simple automatic method to extract robust 3D eye meshes from arbitrary face images and a multi-view consistency regularization which allows to exploit them for improved gaze generalization.
- A revised formulation for gaze estimation, based on dense 3D eye mesh regression from images. To the best of our knowledge, we are the first to utilize an end-to-end 3D eye mesh regression approach for gaze estimation.
- Improved performance over the state-of-the-art in gaze generalization with (10%) and without (23%) using source domain ground truth, with a simple model architecture. Based on that, we believe that 3DGazeNet is an important step towards reliable plug-n-play gaze tracking.

## 2 Related Work

Numerous model designs for supervised 3D gaze estimation have been tested recently, investigating which face region to use as input [12, 42, 82], the model architecture [1,14,46,67] and what external stimuli to utilize to improve performance [52]. Motivated by the difficulties in collecting diverse and large scale data for gaze estimation, recent works have shown that valuable gaze representations can be extracted in fully unsupervised settings, by applying gaze redirection [74] or disentanglement constraints [63].

**Gaze Adaptation and Generalization** Much effort has been made to design methods that adapt well to known target subjects and environments, by employing either few labeled samples [29, 53, 73] or completely unlabeled data of the target domain [5,7,11,24,27,47,68,70]. Differently from the above, gaze generalization models aim to improve cross-domain performance without any knowledge of the target domains. The models in [5, 11, 70], even though targeted for gaze adaptation, are based on learning general features for gaze estimation and thus, they perform well in target domain-agnostic settings. Moreover, [40] has shown

that it is possible to train general gaze estimation models by employing geometric constraints in scenes depicting social interaction between people. We believe that [40] is the closest work to ours, as it is the only method which uses 3D geometric cues of the scene to learn gaze from arbitrary face data. Lastly, [78] proposes to improve generalization by employing synthetic images which are, however, limited by the gaze distribution of existing gaze datasets. Both the implementation and custom dataset are not public, which hinders reproducibility and reliable comparisons.

**Model-Based Gaze Estimation** Differently from the above, sparse or semantic representations of the eye geometry have also been employed by some methods to infer gaze from images [54, 55, 66, 67, 71, 72]. However, such representations do not convey information about the 3D substance of eyes and are prone to noisy predictions. In contrast, by predicting 3D eye meshes we learn a much more robust representation, from which we can retrieve any other sparse or semantic one just by indexing. Recovering dense 3D geometry of the eye region from images by fitting parametric models of the shape and texture has been previously proposed [71]. However, restrictions posed by building large-scale parametric models and fitting in-the-wild images have resulted in low gaze accuracy compared to learning-based methods.

**Face Reenactment and Learning from Synthetic Data** Synthetic image data have been previously used in training deep networks, mainly to augment the training datasets and provide pseudo-ground truth annotations. For instance, [84] used CycleGAN [83] to create a new training corpus in order to balance emotion classes in the task of emotion classification. More recently, GANcraft [28] employed SPADE [56] to generate pseudo-ground truth images that were used to supervise their neural rendering framework. In this work, we obtain access to image pairs of the same subject in different views, by taking advantage of HeadGAN [19], a face reenactment system. In contrast to person-specific reenactment methods [18, 36, 41] or person-generic landmark-driven approaches [69, 75, 76], HeadGAN is able to perform free-view synthesis using a single source image.

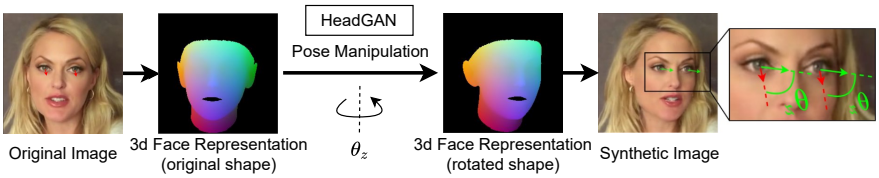



**Fig. 2:** We use HeadGAN [19] to generate novel views by manipulating the 3D pose of the face. During synthesis, angle $\theta_z$ is transferred to all facial parts including the eyes, thus the relative angle between the head and eyes (i.e. the gaze direction in the head coordinate system) is maintained.

# 3 Method

## 3.1 Problem Definition and Motivation

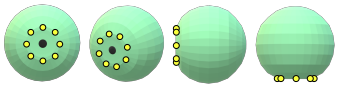
The aim of this work is to design a method that given a face image $\mathbf{I}$, it estimates $2 \times N_v$ 3D coordinates $\mathbf{V} = [\mathbf{V}_l^T, \mathbf{V}_r^T]^T$, where $\mathbf{V}_l \in \mathbb{R}^{N_v \times 3}$ are coordinates corresponding to the left eyeball while $\mathbf{V}_r \in \mathbb{R}^{N_v \times 3}$ to the right, as well as a 3D gaze vector $g = (g_x, g_y, g_z)$. Then, the final gaze result is calculated by the mean direction of the two output components. Inspired by recent work in self-supervised 3D body pose estimation [31, 44, 65], we adopt multi-view constraints to train our model based on in-the-wild faces and automatically generated gaze pseudo-labels.

To employ multi-view losses, we assume that images of the same subject with different head poses and the same gaze direction relatively to the head are available. For example, this condition is satisfied when a face picture is taken from different angles at the same time. As such images are not commonly available for in-the-wild datasets, we employ HeadGAN [19], a recent face reenactment method, to generate novel face poses from existing images. HeadGAN is able to synthesize face animations using dense face geometry, which covers the eyes, as a driving signal and single source images. Using dense geometry guarantees that the relative angle between the head and eyes is maintained when synthesizing novel poses, as it is shown in Fig. 2.

## 3.2 Unified 3D Eye Representation

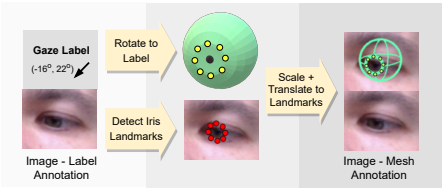
Learning consistent eye meshes across different images and datasets, requires establishing a unified 3D eye representation. To that end, we define a 3D eyeball template as a rigid 3D triangular mesh with spherical shape, consisting of $N_v = 481$ vertices and $N_t = 928$ triangles. We create two mirrored versions, $\mathbf{M}_l$ and $\mathbf{M}_r$, of the above mesh to represent a left and a right reference eyeball respectively. This representation allows us to allocate semantic labels to specific vertices of the eyeball, such as the iris border (Fig. 3 (a)), and calculate 3D gaze direction as the orientation of the central axis of our 3D eyeball template. In practice, an offset angle (the kappa coefficient) exists between the optical (central) and visual axes of eyes, which is subject-dependent and varies between $-2^o$ to $2^o$ across the population [73]. Accounting for this offset is essential for person-specific gaze estimation [29, 45, 53, 73]. However, in our case of cross-dataset and in-the-wild gaze generalization, in which errors are much larger than the possible offset, data diversity is more important than anatomical precision and thus, our spherical eyeball is a reasonable approximation.

**3D Eyes Ground-Truth from Gaze Datasets** For gaze estimation datasets, exact supervision can be acquired by automatically fitting the eyeball template on face images based on sparse iris landmarks and the available gaze labels, as shown in Fig. 3(b). Specifically, we first rotate the eyeball template around its center according to the gaze label. Then, we align (scale and translation) $x$,

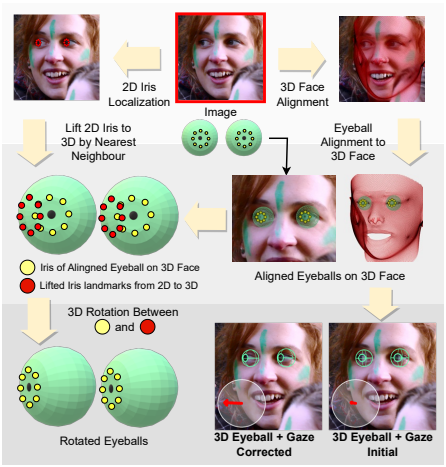


**M**: $N_v = 481$ vertices, $N_t = 928$ triangles

(a) Eyeball template




(b) Ground truth generation




(c) Pseudo-ground truth generation

**Fig. 3:** (a) The employed rigid 3D eyeball mesh template. (b) Ground truth data generation, applied on gaze estimation datasets with available ground truth. (c) Pseudo-ground truth data generation, applied on arbitrary face images without any gaze label.

$y$ coordinates of the rotated eye mesh to the iris landmarks of the image and multiply $z$ coordinates with the same scale. To extract sparse iris landmarks we employed the method of [55] as a basis for building an iris localization model which is robust against occlusions and low resolution. More details about the iris localization model are provided in the supplemental material.

**3D Eyes Pseudo-Ground Truth from In-The-Wild Images** To extract 3D eyes from images without gaze labels, we have developed an automatic pipeline based on 3D face alignment and 2D iris localization. First, we recover the 3D face with $x$, $y$ in image space using an off-the-shelf method. Then, we align our eyeball templates in the eye sockets based on the face's eyelid landmarks and predefined eyelid landmarks around the eyeball templates. In fact, we use the two corner landmarks of each eye which do not move between open and closed eyes. Next, we lift 2D iris predictions to 3D by finding the nearest vertexes from the aligned 3D eye templates. Finally, we compute the rotation between the initially aligned eyes and the 3D-lifted iris center and rotate the eyeballs accordingly. For 3D face alignment, we employ RetinaFace [16] and for 2D iris localization [55] as above. The process is presented in Fig. 3(c).

### 3.3 Joint 3D Eye Mesh and Vector Regression

Given an input face image **I**, we utilize 5 face detection landmarks to crop patches around each one of the two eyes. We resize the patches to shape $128 \times 128 \times 3$ and stack them channel-wise along with a cropped image of the face. We employ a simple model architecture consisting of a ResNet-18 [30] to extract

features, followed by two fully connected layers to map them to two separate eye modalities, which are a) dense 3D eye coordinates and b) a 3D gaze vector. As the final gaze output, we consider the mean direction calculated from the two modalities.

To train the above network for mesh regression, similarly to [16], we enforce a vertex loss and an edge length loss between the model outputs and the respective ground truth or pseudo-ground truth, which can be expressed as:

$$\mathcal{L}_{vert} = \frac{1}{N_v} \sum_{j=\{l,r\}} \sum_{i=1}^{N_v} \|\mathbf{V}_{j,i} - \mathbf{V}^*_{j,i}\|_1, \tag{1}$$

where $\mathbf{V}_j \in \mathbb{R}^{N_v \times 3}$ and $\mathbf{V}^*_j \in \mathbb{R}^{N_v \times 3}$ for $j = \{l, r\}$ are the output and the (pseudo-)ground truth coordinates, while the edge length loss (based on the fixed mesh triangulation of our template meshes) can be written as:

$$\mathcal{L}_{edge} = \frac{1}{3N_t} \sum_{j=\{l,r\}} \sum_{i=1}^{3N_t} \|\mathbf{E}_{j,i} - \mathbf{E}^*_{j,i}\|_2, \tag{2}$$

where $\mathbf{E}_j \in \mathbb{R}^{3N_t}$ and $\mathbf{E}^*_j \in \mathbb{R}^{3N_t}$ for $j = \{l, r\}$ are the edge lengths of the predicted and the (pseudo-)ground truth eyes. As edge length we define the Euclidean distance between two vertices of the same triangle. In addition to the mesh regression losses, we enforce a gaze loss to the gaze output of our model, expressed as:

$$\mathcal{L}_{gaze} = (180/\pi) \arccos(\mathbf{g}^T \mathbf{g}^*), \tag{3}$$

where $\mathbf{g}$ and $\mathbf{g}^*$ are the normalized model output and the gaze (pseudo-)ground truth respectively. We combine losses of Eqs. (1) to (3) in a single loss function to train our models with supervision from (pseudo-)ground truth 3D eye meshes and gaze vectors. The combined loss is written as:

$$\mathcal{L}_{(P)GT} = \lambda_v \mathcal{L}_{vert} + \lambda_e \mathcal{L}_{edge} + \lambda_g \mathcal{L}_{gaze}, \tag{4}$$

where $\lambda_v$, $\lambda_e$, and $\lambda_g$ are parameters which regularize the contribution of the loss terms in the overall loss. From our experiments we have selected their values to be $\lambda_v = 0.1$, $\lambda_e = 0.01$ and $\lambda_g = 1$.

### 3.4 Multi-View Consistency Supervision

Extending our training dataset with in-the-wild images and training using pseudo-ground truth, usually improves the ability of our models to generalize to unseen domains, as can be seen by our experiments in Sec. 4.3. However, automatically generated 3D eyes and gaze include inconsistencies which are hard to identify and filter out. To balance the feedback of direct supervision from pseudo-ground truth, we design a multi-view supervision framework, based on pairs of real and synthetic images with different head poses, generated by HeadGAN as described in Sec. 3.1.

Recovering dense 3D face coordinates and pose from images has recently been quite reliable [2, 16, 16, 23]. Having a pair of images $\mathbf{I}_1$ and $\mathbf{I}_2$ of the same

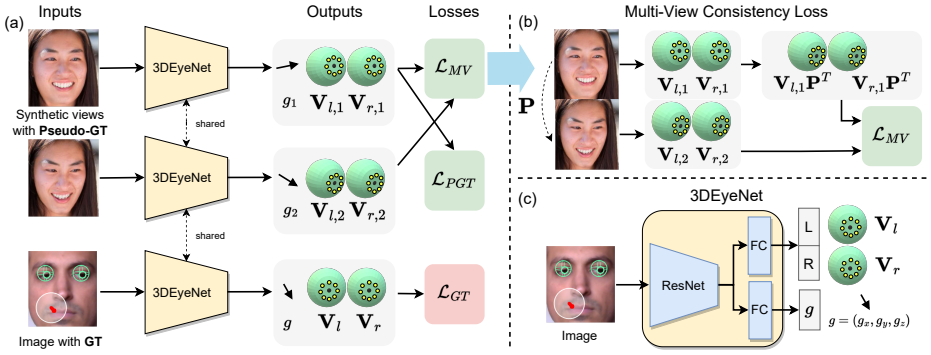



**Fig. 4:** Overview of the proposed method 3DGazeNet. a) During training we employ single images with ground-truth supervision or pairs of synthetic views of the same subject with pseudo-annotations and different head poses. Different sets of losses are employed depending on the type of supervision. b) Detailed demonstration of $\mathcal{L}_{MV}$. 3D transformation $\mathbf{P}$ which maps view 1 to view 2, is employed to transform points $\mathbf{V}_{l,1}$ and $\mathbf{V}_{r,1}$, before calculating an L1 distance loss against $\mathbf{V}_{l,2}$ and $\mathbf{V}_{r,2}$. c) The base network (3DEyeNet) of our model consists of a ResNet-18 backbone and two fully connected layers leading to the 3D eye mesh and gaze vector outputs.

subject and their reconstructed 3D faces, we can compute a transformation matrix $\mathbf{P} \in \mathbb{R}^{3\times4}$ which aligns the two faces in image space. Assuming that gaze direction in both images remains still relative to the face, as is the case with images created by HeadGAN, we are able to supervise 3D regression of eyes by restricting our model's predictions to be consistent over an image pair, as output vertices should coincide when transformation $\mathbf{P}$ is applied to one of the pair's outputs. A similar approach has been employed successfully for weakly-supervised body pose estimation [31,44,65]. Particularly, we form the vertex loss of a pair as:

$$\mathcal{L}_{MV,vertex} = \frac{1}{N_v} \sum_{j=\{l,r\}} \sum_{i=1}^{N_v} \|\mathbf{V}_{1,j,i}\mathbf{P}^T - \mathbf{V}_{2,j,i}\|_1, \tag{5}$$

where $\mathbf{V}_{1,j}, \mathbf{V}_{2,j} \in \mathbb{R}^{N_v\times4}$ for $j = \{l, r\}$ are the output matrices for left and right eyes, which correspond to input images $\mathbf{I}_1$ and $\mathbf{I}_2$. $\mathbf{V}_{1,j,i}, \mathbf{V}_{2,j,i} \in \mathbb{R}^4$ are the specific homogeneous 3D coordinates indexed by $i$ in the above matrices. To enforce consistency constraints to the gaze head of our model, we analyse matrix $\mathbf{P}$ to scale $s$, rotation $\mathbf{R}$ and translation $\mathbf{t}$ components and employ $\mathbf{R}$ in a gaze consistency loss within a pair:

$$\mathcal{L}_{MV,gaze} = (180/\pi)\arccos((\mathbf{g}_1^T\mathbf{R}^T)\mathbf{g}_2), \tag{6}$$

where $\mathbf{g}_1$ and $\mathbf{g}_2$ are the normalized model outputs for input images $\mathbf{I}_1$ and $\mathbf{I}_2$ respectively. We combine losses of Eqs. (5) and (6) in a single loss function to enforce multi-view consistency in mesh and gaze vector regression, between

model outputs coming from pairs of input images. This loss is written as:

$$\mathcal{L}_{MV} = \lambda_{MV,v}\mathcal{L}_{MV,vertex} + \lambda_{MV,g}\mathcal{L}_{MV,gaze}, \tag{7}$$

where $\lambda_{MV,v}$ and $\lambda_{MV,g}$ are parameters which regularize the contribution of the loss terms in the overall loss. In our experiments, we have selected their values to be $\lambda_{MV,v} = 0.1$ and $\lambda_{MV,g} = 1$. To train models with all supervision signals, i.e. ground truth ($\mathcal{L}_{GT}$), pseudo-ground truth ($\mathcal{L}_{PGT}$) and multi-view supervision ($\mathcal{L}_{MV}$), we utilize the following overall loss function:

$$\mathcal{L} = \lambda_{GT}\mathcal{L}_{GT} + \lambda_{PGT}\mathcal{L}_{PGT} + \lambda_{MV}\mathcal{L}_{MV}, \tag{8}$$

with parameters $\lambda_{GT} = \lambda_{PGT} = \lambda_{MV} = 1$. Implementation details are included in the supplemental material. An overview of 3DGazeNet is presented in Fig. 4.

## 4 Experiments

### 4.1 Datasets

**Gaze Datasets** Captured in a lab environment, ETH-XGaze (EXG) [79] consists of 756K frames of 80 subjects and includes large head pose and gaze variation. Collected in uncontrolled indoor environments with mobile devices, MPIIFaceGaze (MPII) [81] includes smaller head pose and gaze variation and consists of 45K images of 15 subjects, while GazeCapture (GC) [42] contains almost 2M frontal face images of 1474 subjects. In contrast to the above datasets, Gaze360 (G360) [35] is the only gaze dataset captured both indoors and outdoors and consists of 127K training sequences from 365 subjects. The large variation in head pose, gaze, and environmental conditions of Gaze360 makes it the most challenging yet appropriate benchmark for in-the-wild gaze estimation, available in literature. For our experiments, we normalized the above datasets based on [80], except for Gaze360 which we process to get normalized face crops. Additionally, we employ the predefined training-test splits, while for Gaze360 we only use the frontal facing images with head pose yaw angle up to $90^o$. The head pose and gaze distributions of the above datasets are presented in Fig. 5.

**In-The-Wild Face Datasets** In-the-wild face datasets consist of significantly more unique subjects and capturing environments. For our experiments, we employed four publicly-available datasets FFHQ [33] (70K images), AFLW [38] (25K images), AVA [25, 48, 49] and CMU-Panoptic [32]. FFHQ and AFLW are in-the-wild face datasets commonly used for face analysis, AVA is a large-scale in-the-wild human activity dataset annotated under the Looking-At-Each-Other condition and CMU-Panoptic is collected in lab conditions and captures interactions of multiple people in the same scene. FFHQ and AFLW include one face per image and thus are only processed to get normalized face crops. AVA and CMU-Panoptic include frames with multiple faces, from which we randomly select 80K faces from each dataset with a maximum head pose of $90^o$. Similarly to [40], for CMU we employed only frames captured with cameras in eye height. We name this collection of 255K images as the "In-The-Wild Gaze" dataset (ITWG). Lastly, to enforce multi-view supervision as described in Sec. 3.4, we synthesized

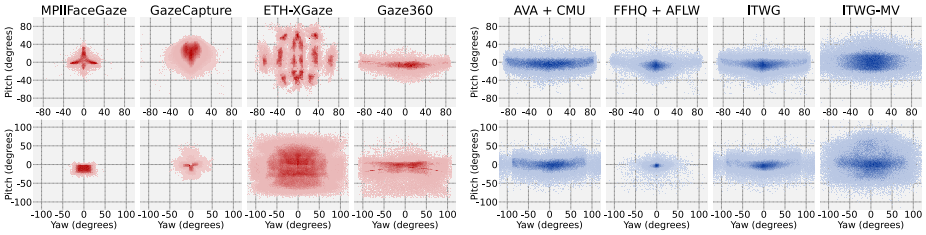



**Fig. 5:** Distributions of the head pose (top row) and gaze (bottom row) of the gaze datasets (red) and the face datasets (blue). Wide distribution datasets CMU, AVA, FFHQ, and AFLW are exploited to close the gap between diverse image domains.

novel views from images of ITWG using HeadGAN, sampling the pitch and yaw angles from Gaussians $\mathcal{N}(0, 20)$, relatively to the original head pose. We name this collection of images as "Multi-View In-The-Wild Gaze" dataset (ITWG-MV) and employ it to improve the generalization of gaze estimation. The head pose and gaze distributions of the above datasets are presented in Fig. 5.

### 4.2 Gaze Generalization

In this section, we evaluate 3DGazeNet in within-dataset and cross-dataset experiments. We believe that [40] is the most closely related method to ours, as it is the only method using 3D geometric cues of the scene to generalize gaze from arbitrary face data.

**Cross-dataset Evaluation** We design two cross-dataset experiments to test the generalization of our method on G360 and report the results on Tab. 1(a) and (b). Particularly, the experiments are: a) we train our method on the CMU, AVA, and ITWG-MV datasets utilizing only our pseudo-labels and multi-view supervision and b) we additionally employ ground truth supervision from GC and EXG. From the results of the above experiments, it becomes obvious that our geometry-aware pseudo-labels employed within our multi-view supervision training effectively generalize gaze estimation to unseen domains, even without any available ground truth. In particular, in experiment a) our method outperforms [40] by 23% with AVA, 22% with CMU, 12.5% with AVA+CMU and 20% with our large-scale ITWG-MV. Similarly, in experiment b) 3DGazeNet outperforms [40] by 10% and 9% with GC and EXG respectively.

**Within-dataset Evaluation** Here we compare our method against state-of-the-art within-dataset gaze estimation on G360. Similarly to [40], we employ AVA for additional supervision, while we also examine the effect of the larger-scale ITWG-MV. The results, presented in Tab. 1 (c), show that multi-view supervision from AVA does not improve performance (which is in line with the compared method), but the large-scale ITWG-MV does.

**Comparison with state-of-the-art** We further compare 3DGazeNet against recent methods for gaze generalization. The works in [5, 70] are developed with a

**Table 1:** Weakly-supervised method evaluation in cross- and within-dataset experiments. In all cases, we calculate gaze error in degrees (lower is better), on the test set of Gaze360. CMU and AVA correspond to subsets of ITWG-MV (i.e. augmented for multi-view supervision), providing a clearer comparison with [40]. Our method trained with ITWG-MV outperforms the baselines in all cases. 3DGN referes to 3DGazeNet

**(a) Cross-dataset Synthetic Views**

| Dataset | [40] | 3DGN |
|---|---|---|
| AVA | 29.0 | **22.4** |
| CMU | 26.0 | **20.3** |
| CMU+AVA | 22.5 | **19.7** |
| ITWG-MV | - | **18.1** |

**(b) Cross-dataset Ground Truth + Synthetic Views**

| Dataset | [79] | [40] | 3DGN |
|---|---|---|---|
| GC | 30.2 | 29.2 | **27.5** |
| GC+AVA | - | 19.5 | **18.9** |
| GC+AVA+CMU | - | - | **18.4** |
| GC+ITWG-MV | - | - | **17.6** |
| EXG | 27.3 | 20.5 | 22.1 |
| EXG+AVA | - | 16.9 | 17.1 |
| EXG+AVA+CMU | - | - | **16.7** |
| EXG+ITWG-MV | - | - | **15.4** |

**(c) Within-dataset Ground Truth + Synthetic Views**

| Dataset | [35] | [40] | 3DGN |
|---|---|---|---|
| G360 | 11.1 | 10.1 | **9.6** |
| G360+AVA | - | 10.2 | **9.7** |
| G360+AVA+CMU | - | - | **9.5** |
| G360+ITWG-MV | - | - | **9.3** |

**Table 2:** Comparison with state-of-the-art in domain generalization for gaze estimation. In all experiments our model outperforms the compared methods. Gaze error is in degrees (lower is better).

| | Stage 1 (Gaze Generalization Models) | | | | | | | | + Stage 2 (Adaptation/Fine Tuning) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EXG | | EXG+ITWG-MV | | G360 | | G360+ITWG-MV | | EXG+ITWG-MV | | G360+ITWG-MV | |
| Method | MPII | GC | MPII | GC | MPII | GC | MPII | GC | MPII | GC | MPII | GC |
| RAT/RUDA [5] | 7.1 | **8.4** | 7.0 | 8.2 | 9.3 | 9.0 | 9.1 | 8.5 | 6.8 | 8.1 | 7.9 | 8.3 |
| CDG/CRGA [70] | **6.7** | 9.2 | 6.9 | 9.5 | **7.0** | **8.3** | 8.1 | 8.9 | 7.4 | 9.0 | 7.6 | 8.7 |
| PureGaze [11] | 7.9 | 8.7 | 7.7 | 9.3 | 7.6 | 8.3 | 7.4 | 8.6 | 6.6 | 8.0 | 7.2 | 8.3 |
| 3DGazeNet | 7.7 | 10.7 | **6.0** | **7.8** | 9.1 | 12.1 | **6.3** | **8.0** | - | - | - | - |

focus on domain adaptation for gaze estimation and encompass two-stage training schemes, both training feature invariant models at the first stage. That is, in the first training stage RUDA [5] trains gaze estimation model invariant to image rotations, while CRGA [70] uses a contrastive loss to separate image features according to gaze. The second stage of the above methods is focused on adapting the initially trained models to specific target domains. As our method aims to train general gaze estimation models without knowledge of specific target domains, we implement the first-stage models of the above methods, namely RAT [5], CDG [70] and compare them with 3DGazeNet in cross-dataset experiments. Additionally, we compare against PureGaze [11] which is a gaze generalization method that purifies face features to achieve higher gaze estimation performance. To follow the evaluation protocol in the above works, we train all methods on EXG and G360 (+ITWG-MV) and test on MPII and GC. For completeness, we include results of the full models RUDA and CRGA after using ITWG-MV according to their domain adaptation schemes. For PureGaze, ITWG-MV was used for fine-tuning. Tab. 2 shows that the proposed method outperforms the baselines for gaze generalization when ITWG-MV is employed. The compared methods do not include regularization for the noisy labels of ITWG-MV, resulting in similar or worse performance, while our method exploits them through $\mathcal{L}_{MV}$, benefiting from the extended variation.

**Table 3:** Comparison between training targets Vector(V), Mesh(M) and Mesh+Vector(M+V) in within-dataset experiments (using only $\mathcal{L}_{GT}$). Target M+V leads to lower errors than state-of-the-art. Error is in degrees (lower is better).

| Dataset | Compared Methods | | | | | | | | 3DGazeNet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | [51] | [13] | [1] | [15, 82] | [53] | [15, 35] | [40] | [79] | V | M | M+V |
| MPII | 4.04 | 4.00 | **3.92** | 4.9 | 5.3 | 4.06 | - | 4.8 | 4.1 | 4.2 | 4.0 |
| G360 | 10.7 | 10.6 | 10.4 | 14.9 | - | 11.1 | 10.1 | - | 9.8 | 9.8 | **9.6** |
| GC | - | - | - | - | 3.5 | - | - | 3.3 | 3.2 | 3.3 | **3.1** |
| EXG | - | - | - | 7.3 | - | - | - | 4.5 | **4.2** | 4.4 | **4.2** |

### 4.3 Ablation studies

**Gaze Estimation via 3D Eye Mesh Regression** Here we experimentally evaluate our suggestion that gaze estimation benefits from replacing the training target from gaze vectors or angles to dense 3D eye coordinates. To this end, we employ the fully supervised version of our model, utilizing data with exact ground truth and $\mathcal{L}_{GT}$ for training. We conduct within-dataset experiments on MPII, GC, G360 and EXG for which specific training-testing subsets are provided. We compare against state-of-the-art methods [1, 13, 15, 35, 40, 51, 53, 79, 82] and report the results in Tab. 3. In almost all cases, our model outperforms the baselines, while combining the two modalities, i.e. dense 3D meshes and gaze vectors (M+V), improves performance compared to training with vector targets (V) or 3D mesh targets (M) alone. This is possibly due to the distinct nature of the two modalities, i.e. the vectors provide exact label supervision, while meshes provide a robust representation which limits sparse prediction errors.

The main benefit of dense coordinate regression over pose parameters or sparse points prediction is that individual parameter errors have limited effect on the total outcome making them more robust to prediction inaccuracies [16]. This effect is particularly useful for our multi-view training scheme in which introducing consistency of dense correspondences between images rather than only vector consistency, offers stronger regularization. We validate this argument in gaze generalization experiments in G360, GC, EXG, and MPII, presented in Tab. 4. For this experiment, we consider three versions of 3DGazeNet: one which predicts only gaze vectors and no coordinates (Vector), one which predicts 8 3D iris landmarks instead of dense eye meshes (Iris+Vector), to highlight the effect of dense coordinate prediction, and the full 3DGazeNet (Mesh+Vector). The results show that employing combined training targets always benefits performance, while replacing dense 3D eye meshes with iris landmarks highly limits this effect.

**The Effect of Gaze Pseudo-Labels and Multi-View Supervision** Here we examine the contribution of our automatic geometry-aware pseudo-labels and the multi-view supervision loss of our approach. To this end, we consider three training scenarios which are the following: a) training with ITWG and its pseudo-labels as ground truth ($\mathcal{L}_{PGT}$), b) training with ITWG-MV utilizing only the multi-view consistency constraints and no pseudo-labels ($\mathcal{L}_{MV}$) and c) training with ITWG-MV while employing both pseudo-labels and the multi-view

**Table 4:** Comparison between training targets Vector, Iris+Vector and Mesh+Vector for domain generalization when employing our full model (Eq. (8)). For the target Vector, we remove all mesh terms from the employed losses. In all experiments, the target Mesh+Vector results in a lower error. Gaze error is in degrees (lower is better).

| | Vector | | | | Iris+Vector | | | | Mesh+Vector | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Training Dataset | G360 | GC | EXG | MPII | G360 | GC | EXG | MPII | G360 | GC | EXG | MPII |
| ITWG-MV | 19.1 | 10.1 | **16.7** | 8.5 | 18.8 | 9.9 | **16.7** | 8.2 | **18.1** | **9.0** | **16.7** | **7.6** |
| G360+ITWG-MV | 10.1 | 10.2 | 15.1 | 7.0 | 9.7 | 9.4 | 15.0 | 6.8 | **9.3** | **8.0** | **14.6** | **6.3** |
| GC+ITWG-MV | 18.2 | 3.1 | 16.0 | **6.1** | 18.0 | **3.0** | 15.9 | 6.2 | **17.6** | **3.0** | **15.5** | **6.1** |
| EXG+ITWG-MV | 16.5 | 10.2 | 4.5 | 6.6 | 16.3 | 9.6 | 4.5 | 6.4 | **15.4** | **7.8** | **4.3** | **6.0** |
| MPII+ITWG-MV | 17.8 | 8.2 | 15.2 | 4.8 | 17.9 | 7.6 | 15.0 | 4.6 | **17.6** | **6.8** | **14.9** | **4.2** |

**Table 5:** The effect of incorporating pseudo-ground truth and multi-view supervision during training. Both components contribute towards improving results in cross-dataset gaze estimation experiments. Gaze error is in degrees (lower is better).

| Dataset | $\mathcal{L}_{GT}$ | $\mathcal{L}_{PGT}$ | $\mathcal{L}_{MV}$ | G360 | GC | EXG | MPII |
|---|---|---|---|---|---|---|---|
| ITWG | - | ✓ | - | 23.1 | 14.8 | 24.3 | 13.6 |
| ITWG-MV | - | - | ✓ | 47.4 | 33.2 | 41.1 | 32.8 |
| ITWG-MV | - | ✓ | ✓ | **18.1** | **9.0** | **16.7** | **7.6** |
| GC | ✓ | - | - | 27.5 | 3.1 | 28.4 | 10.4 |
| GC+ITWG | ✓ | ✓ | - | 21.4 | 3.2 | 23.7 | 9.1 |
| GC+ITWG-MV | ✓ | - | ✓ | 24.7 | 3.5 | 26.2 | 10.1 |
| GC+ITWG-MV | ✓ | ✓ | ✓ | **17.6** | **3.0** | **15.5** | **6.1** |

consistency loss ($\mathcal{L}_{PGT}$+$\mathcal{L}_{MV}$). To further evaluate the effect of the pseudo-labels and multi-view loss, we repeat the above experiments by adding ground truth supervision from GC (+$\mathcal{L}_{GT}$). We test our models on the test set of G360, GC, EXG, and MPII, and report the results in Tab. 5. In all cases, combining our pseudo-labels and multi-view loss yields the lowest error in degrees. Lastly, utilizing only $\mathcal{L}_{MV}$ on ITWG-MV leads to very high errors which is reasonable as no supervision for the eyeball topology exists, thus, the model outputs cannot follow the spherical shape of the eyeball template.

**The Effect of Head Pose Distribution of ITWG** Head pose distribution difference between the train and test set is one of the main reasons that gaze-estimation models fail in cross-dataset situations. To close the gap between different training and testing scenarios, we have designed ITWG, a large-scale dataset with widespread variation in head pose and gaze angles. To study the effect of the head pose variation of ITWG in our experiments, we employ different subsets of ITWG with various levels of head pose variation and conduct cross-dataset experiments with them. In particular, we consider four subsets of ITWG, with maximum yaw angles of $5^o$, $20^o$, $40^o$ and $90^o$ (all) respectively.

We train 3DGazeNet with ground truth supervision from MPII as well as pseudo-labels and multi-view supervision from the four versions of ITWG-MV.

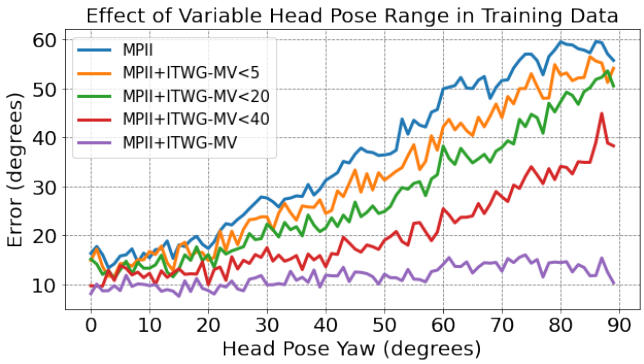



**Fig. 6:** Gaze error of G360 across head poses when training with MPII and subsets of ITWG-MV. Wider range of head poses in the ITWG-MV data, lead to significantly lower errors in large poses.

The results of testing on G360 are presented in Fig. 6. The resulting curves clearly demonstrate the effect of the available head pose variation in the training data. Specifically, utilizing the entirety of ITWG-MV leads to the lowest errors which are relatively consistent across the head pose range. As expected, decreasing the available head pose variation, increasingly affects model performance with the worst case being training with MPII alone. Based on the above finding we argue that the gap between small and wide distribution gaze datasets (regarding head pose) can effectively close by employing similarly large distribution unlabeled face datasets, which is crucial for training plug-n-play gaze estimation models that can be directly employed in applications.

## 5 Limitations and Conclusion

In Sec. 4, we shown that pseudo-ground truth can be effectively utilized in gaze estimation. Nevertheless, a limitation of our method is that pseudo-annotation accuracy is related to the accuracy of 3D face and 2D iris alignment. In addition, our current method cannot operate on images without a visible face (when the face is looking away from the camera).

Overall, In this work, we present a novel weakly-supervised method for gaze generalization, based on dense 3D eye mesh regression. We demonstrate that by utilizing both 3D eye coordinates and gaze labels during training, instead of just gaze labels, we can achieve lower prediction errors. Moreover, we explore the possibility of exploiting the abundantly available in-the-wild face data for improving gaze estimation generalization. To that end, we propose a novel methodology to generate robust, 3D geometry-aware pseudo ground truth labels, as well as a multi-view weak-supervision framework for effective training. By enforcing these constraints, we are able to successfully utilize in-the-wild face data and achieve improvements in cross-dataset and within-dataset experiments.

**Acknowledgments.** S. Zafeiriou was supported by EPSRC Project DEFORM (EP/S010203/1) and GNOMON (EP/X011364).

## References

1. Abdelrahman, A.A., Hempel, T., Khalifa, A., Al-Hamadi, A., Dinges, L.: L2cs-net: Fine-grained gaze estimation in unconstrained environments. In: ICFSP. pp. 98–102. IEEE (2023)
2. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: Face alignment and detection via 6dof, face pose estimation. In: CVPR (2021)
3. Alp Guler, R., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I.: Densereg: Fully convolutional dense shape regression in-the-wild. In: CVPR (2017)
4. Andrist, S., Tan, X.Z., Gleicher, M., Mutlu, B.: Conversational gaze aversion for humanlike robots. In: HRI (2014)
5. Bao, Y., Liu, Y., Wang, H., Lu, F.: Generalizing gaze estimation with rotation consistency. In: CVPR (2022)
6. Burova, A., Mäkelä, J., Hakulinen, J., Keskinen, T., Heinonen, H., Siltanen, S., Turunen, M.: Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In: CHI (2020)
7. Cai, X., Zeng, J., Shan, S., Chen, X.: Source-free adaptive gaze estimation by uncertainty reduction. In: CVPR. pp. 22035–22045 (2023)
8. Cai, Y., Ge, L., Cai, J., Yuan, J.: Weakly-supervised 3d hand pose estimation from monocular rgb images. In: ECCV (2018)
9. Castner, N., Kuebler, T.C., Scheiter, K., Richter, J., Eder, T., Hüttig, F., Keutel, C., Kasneci, E.: Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In: ACM ETRA (2020)
10. Chen, M., Jin, Y., Goodall, T., Yu, X., Bovik, A.C.: Study of 3d virtual reality picture quality. IEEE Journal of Selected Topics in Signal Processing (2020)
11. Cheng, Y., Bao, Y., Lu, F.: Puregaze: Purifying gaze feature for generalizable gaze estimation. In: AAAI (2022)
12. Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: AAAI (2020)
13. Cheng, Y., Lu, F.: Gaze estimation using transformer. In: ICPR (2022)
14. Cheng, Y., Lu, F., Zhang, X.: Appearance-based gaze estimation via evaluation-guided asymmetric regression. In: ECCV (2018)
15. Cheng, Y., Wang, H., Bao, Y., Lu, F.: Appearance-based gaze estimation with deep learning: A review and benchmark. arXiv preprint arXiv:2104.12668 (2021)
16. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: CVPR (2020)
17. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: CVPR Workshops (2019)
18. Doukas, M.C., Koujan, M.R., Sharmanska, V., Roussos, A., Zafeiriou, S.: Head2head++: Deep facial attributes re-targeting. T-BIOM (2021)
19. Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: One-shot neural head synthesis and editing. In: ICCV (2021)
20. Fischer, T., Chang, H.J., Demiris, Y.: Rt-gene: Real-time eye gaze estimation in natural environments. In: ECCV (2018)

21. Fuhl, W., Kasneci, G., Kasneci, E.: Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. ISMAR (2021)
22. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: ACM ETRA (2014)
23. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: CVPR (2019)
24. Ghosh, S., Hayat, M., Dhall, A., Knibbe, J.: Mtgls: Multi-task gaze estimation with limited supervision. In: WACV (2022)
25. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018)
26. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: CVPR (2019)
27. Guo, Z., Yuan, Z., Zhang, C., Chi, W., Ling, Y., Zhang, S.: Domain adaptation gaze estimation by embedding with prediction consistency. In: ACCV (2020)
28. Hao, Z., Mallya, A., Belongie, S., Liu, M.Y.: GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In: ICCV (2021)
29. He, J., Pham, K., Valliappan, N., Xu, P., Roberts, C., Lagun, D., Navalpakkam, V.: On-device few-shot personalization for real-time gaze estimation. In: ICCV Workshops (2019)
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
31. Iqbal, U., Molchanov, P., Kautz, J.: Weakly-supervised 3d human pose learning via multi-view images in the wild. In: CVPR (2020)
32. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV (2015)
33. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
34. Kasahara, I., Stent, S., Park, H.S.: Look both ways: Self-supervising driver gaze estimation and road scene saliency. In: ECCV (2022)
35. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., , Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: ICCV (2019)
36. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollöfer, M., Theobalt, C.: Deep video portraits. TOG (2018)
37. Kleinke, C.L.: Gaze and eye contact: a research review. Psychological bulletin (1986)
38. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: ICCVW (2011)
39. Konrad, R., Angelopoulos, A., Wetzstein, G.: Gaze-contingent ocular parallax rendering for virtual reality. In: TOG (2019)
40. Kothari, R., De Mello, S., Iqbal, U., Byeon, W., Park, S., Kautz, J.: Weakly-supervised physically unconstrained gaze estimation. In: CVPR (2021)
41. Koujan, M.R., Doukas, M.C., Roussos, A., Zafeiriou, S.: Head2head: Video-based neural head synthesis. In: FG (2020)
42. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: CVPR (2016)

43. Kulon, D., Guler, R.A., Kokkinos, I., Bronstein, M.M., Zafeiriou, S.: Weakly-supervised mesh-convolutional hand reconstruction in the wild. In: CVPR (2020)
44. Li, Y., Li, K., Jiang, S., Zhang, Z., Huang, C., Xu, R.Y.D.: Geometry-driven self-supervised method for 3d human pose estimation. In: AAAI (2020)
45. Liu, G., Yu, Y., Mora, K., Odobez, J.: A differential approach for gaze estimation with calibration. In: BMVC (2018)
46. Liu, G., Yu, Y., Mora, K.A.F., Odobez, J.M.: A differential approach for gaze estimation with calibration. In: BMVC (2018)
47. Liu, Y., Liu, R., Wang, H., Lu, F.: Generalizing gaze estimation with outlier-guided collaborative adaptation. In: ICCV (2021)
48. Marín-Jiménez, M.J., Kalogeiton, V., Medina-Suárez, P., , Zisserman, A.: LAEO-Net++: revisiting people Looking At Each Other in videos. TPAMI (2021)
49. Marin-Jimenez, M.J., Kalogeiton, V., Medina-Suarez, P., Zisserman, A.: Laeo-net: Revisiting people looking at each other in videos. In: CVPR (2019)
50. Mavely, A.G., Judith, J.E., Sahal, P.A., Kuruvilla, S.A.: Eye gaze tracking based driver monitoring system. In: ICCS (2017)
51. O Oh, J., Chang, H.J., Choi, S.I.: Self-attention with convolution and deconvolution for efficient eye gaze estimation from a full face image. In: CVPRW (2022)
52. Park, S., Aksan, E., Zhang, X., Hilliges, O.: Towards end-to-end video-based eye-tracking. In: ECCV (2020)
53. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: ICCV (2019)
54. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: ECCV (2018)
55. Park, S., Zhang, X., Bulling, A., Hilliges, O.: Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In: ACM ETRA (2018)
56. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)
57. Richard, A., Lea, C., Ma, S., Gall, J., de la Torre, F., Sheikh, Y.: Audio- and gaze-driven facial animation of codec avatars. In: WACV (2021)
58. Riza Alp Guler, Natalia Neverova, I.K.: Densepose: Dense human pose estimation in the wild. In: CVPR (2018)
59. Smith, B., Yin, Q., Feiner, S., Nayar, S.: Gaze Locking: Passive Eye Contact Detection for Human?Object Interaction. In: ACM UIST (2013)
60. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: CVPR (2014)
61. Sun, J., Wang, X., Shi, Y., Wang, L., Wang, J., Liu, Y.: Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. ACM TOG **41**(6), 1–10 (2022)
62. Sun, J., Wang, X., Wang, L., Li, X., Zhang, Y., Zhang, H., Liu, Y.: Next3d: Generative neural texture rasterization for 3d-aware head avatars. In: CVPR (2023)
63. Sun, Y., Zeng, J., Shan, S., Chen, X.: Cross-encoder for unsupervised gaze representation learning. In: ICCV (2021)
64. Vidal, M., Turner, J., Bulling, A., Gellersen, H.: Wearable eye tracking for mental health monitoring. Computer Communications (2012)
65. Wandt, B., Rudolph, M., Zell, P., Rhodin, H., Rosenhahn, B.: Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In: CVPR (2021)
66. Wang, K., Ji, Q.: Real time eye gaze tracking with 3d deformable eye-face model. In: ICCV (2017)
67. Wang, K., Zhao, R., Ji, Q.: A hierarchical generative model for eye image synthesis and eye gaze estimation. In: CVPR (2018)

68. Wang, K., Zhao, R., Su, H., Ji, Q.: Generalizing eye tracking with bayesian adversarial learning. In: CVPR (2019)
69. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: NeurIPS (2019)
70. Wang, Y., Jiang, Y., Li, J., Ni, B., Dai, W., Li, C., Xiong, H., Li, T.: Contrastive regression for domain adaptation on gaze estimation. In: CVPR (2022)
71. Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: A 3d morphable eye region model for gaze estimation. In: ECCV (2016)
72. Yu, Y., Liu, G., Odobez, J.M.: Deep multitask gaze estimation with a constrained landmark-gaze model. In: ECCV Workshops (2018)
73. Yu, Y., Liu, G., Odobez, J.M.: Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In: CVPR (2019)
74. Yu, Y., Odobez, J.M.: Unsupervised representation learning for gaze estimation. In: CVPR (2020)
75. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. ICCV (2019)
76. Zakharov, E., Ivakhnenko, A., Shysheya, A., Lempitsky, V.: Fast bi-layer neural synthesis of one-shot realistic head avatars. In: ECCV (2020)
77. Zhang, J., Chen, J., Tang, H., Wang, W., Yan, Y., Sangineto, E., Sebe, N.: Dual in-painting model for unsupervised gaze correction and animation in the wild. In: ACM MM (2020)
78. Zhang, M., Liu, Y., Lu, F.: Gazeonce: Real-time multi-person gaze estimation. In: CVPR (2022)
79. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: ECCV (2020)
80. Zhang, X., Sugano, Y., Bulling, A.: Revisiting data normalization for appearance-based gaze estimation. In: ACM ETRA (2018)
81. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: CVPR (2015)
82. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: CVPRW (2017)
83. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
84. Zhu, X., Liu, Y., Li, J., Wan, T., Qin, Z.: Emotion classification with data augmentation using generative adversarial networks. In: PAKDD (2018)