Weakly-Supervised Gaze Estimation from Synthetic Views

Evangelos Ververas1 Polydefkis Gkagkos1 Jiankang Deng1 Jia Guo3
Michail Christos Doukas1,2 Stefanos Zafeiriou1,2
1Huawei Technologies, London, UK 2Imperial College London, UK 3InsightFace
{evangelos.ververas, polydefkis.gkagkos, jiankang.deng, michail.christos.doukas, stefanos.zafeiriou}@huawei.com guojia@gmail.com

Abstract

3D gaze estimation is most often tackled as learning a direct mapping between input images and the gaze vector or its spherical coordinates. Recently, it has been shown that pose estimation of the face, body and hands benefits from revising the learning target from few pose parameters to dense 3D coordinates. In this work, we leverage this observation and propose to tackle 3D gaze estimation as regression of 3D eye meshes. We overcome the absence of compatible ground truth by fitting a rigid 3D eyeball template on existing gaze datasets and propose to improve generalization by making use of widely available in-the-wild face images. To this end, we propose an automatic pipeline to retrieve robust gaze pseudo-labels from arbitrary face images and design a multi-view supervision framework to balance their effect during training. In our experiments, our method achieves improvement of 30% compared to state-of-the-art in cross-dataset gaze estimation, when no ground truth data are available for training, and 7% when they are. We make our project publicly available at https://github.com/Vagver/dense3Deyes

1. Introduction

Eye gaze serves as a cue for understanding human behavior and intents, including attention, communication and mental state. As a consequence, gaze information has been exploited by a lot of applications of various fields of interest, ranging from medical and psychological analysis [6, 32, 56] to human-computer interaction [2], efficient rendering in VR/AR headset systems [4, 7, 34], virtual character animation [51, 68] and driver state monitoring [29, 44].

Several methods have been proposed recently to improve automatic 3D gaze estimation from monocular RGB images. Typically, these methods predict gaze as either a 3D vector or spherical coordinates indicating the direction that someone is looking at, without considering any geometric representation of the eyes. Nevertheless, it has been shown that unconstrained face and body pose estimation from single images benefit from replacing predicting few pose or model parameters by directly predicting dense 3D geometry [12, 21, 38, 52]. To our knowledge, this observation has
not been leveraged for eyes, and thus recovering gaze as a by-product of dense 3D eye mesh reconstruction remains open for investigation.

Training to predict 3D geometry from images requires supervision from related ground truth. In [17] the authors have proposed a dataset of IR images and 3D eyes parameterized by the radius and eye center. However, IR images cannot be directly employed for RGB based methods. In addition, several gaze datasets have become recently available [16, 18, 30, 37, 46, 53, 54, 70, 71]. A straightforward approach to obtain 3D ground truth for these data, is to fit an eyeball using sparse eye landmarks and the available gaze labels. Still, collecting gaze datasets is a costly and challenging process which restricts them being captured in controlled environments and often consisting of limited different identities. This causes the most common challenge in gaze estimation, which is cross-domain generalization. Nevertheless, images and videos of people in-the-wild are abundantly available in the internet. Thus, a reasonable question would be: “Is it possible to utilize in-the-wild face images for improving generalization of gaze-estimation models?”.

In this work, we propose to tackle gaze estimation as end-to-end regression of dense 3D eye coordinates. We acquire compatible 3D ground truth by defining a unified eye representation for all employed datasets, i.e. a rigid 3D eyeball template (Fig. 3 (a)), which we fit on existing gaze datasets based on sparse landmarks and the available gaze labels. Furthermore, we tackle the challenge of cross-domain generalization by taking advantage of largely available in-the-wild face data and recent advances in weak-supervision of training CNNs for human perception tasks [5, 13, 26, 39, 57]. An overview of our method is presented in Fig. 1.

To obtain viable supervision from arbitrary face data, we propose a pipeline to automatically generate robust 3D pseudo-ground truth without any prior gaze information. To balance errors coming from automatic 3D eye mesh generation, we implement a weak-supervision, multi-view constraint which encourages our model to maintain consistency between the 3D eyes across multiple synthetic views of the same subject. We acquire novel views of a face by employing HeadGAN [15], a recently proposed method for face reenactment, which enables us to animate single images. HeadGAN manages to synthesize novel head poses while maintaining the relative difference between the gaze direction and head orientation in the generated image. This happens because in HeadGAN image synthesis is conditioned on dense 3D representations of the face, which includes the eye regions.

We evaluate our method on common gaze estimation datasets including the in-the-wild Gaze360 [30], which is the most challenging yet appropriate benchmark in literature for in-the-wild gaze estimation. We demonstrate that adding dense 3D eye coordinates in the training targets of a gaze estimation system, leads to improved performance in within-dataset experiments. Additionally, we showcase that learning meaningful gaze information from in-the-wild face images, is possible by utilizing pseudo-ground truth within a multi-view training approach, even when no gaze ground truth is available. In this way, we are able to improve gaze estimation generalization to unseen domains and close the gap between common gaze datasets and in-the-wild, real-world environments.

To summarize, the key contributions of our work are:

- We revise the common approach of tackling 3D gaze estimation and propose to learn gaze as a by-product of dense regression of 3D eye meshes from images. To the best of our knowledge, we are the first to adopt an end-to-end dense 3D eye regression approach to support gaze estimation.
- We design an automatic method to generate robust pseudo-ground truth of eyes and gaze from arbitrary face images and utilize them for improving gaze estimation generalization to unseen domains.
- We propose a weakly-supervised framework to balance the effect of inaccurate pseudo-ground truth during training. In particular, we employ synthetic views of face images and employ multi-view consistency constraints, which allow us to further improve gaze model generalization.

2. Related Work

3D Gaze Estimation Undoubtedly, the most common approach to tackle gaze estimation has been by learning a direct mapping between eye or face images and few gaze coordinates or angles. To this end, numerous model design settings have been investigated recently, including the face region to use as input [8, 37, 73], the model architecture [10, 40, 58] and what external stimuli to utilize to improve performance [46]. Nevertheless, much effort has also been made to design domain adaptation methods, to generalize well to unseen subjects and environments, by employing either few labeled samples [24, 47, 64] or completely unlabeled data of the target domain [3, 22, 41, 59, 61].

Better yet, in recent works it has been shown that learning gaze from images can be achieved in fully unsupervised settings. Particularly, valuable gaze representations can be extracted from image encoder-decoder architectures by applying gaze redirection [65] or disentanglement [55] constraints. In addition, [35] shown that it is possible to train gaze estimation by employing geometric constraints in scenes depicting social interaction and particularly scenes of people looking at each other (LAEO). We believe that [35] is the closest work to ours, as it is the
only method which utilizes arbitrary face-data to learn gaze information. Lastly, even though [69] proposes a method in the same line of research, the presented experiments are conducted on their proposed dataset MPSGaze which has not become available yet.

Differently from the above, sparse or semantic representations of the eye geometry have also been employed by some methods to infer gaze from images [48,49,58,62,63]. 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 are able to learn a much more robust representation, from which we can retrieve any other sparse or semantic one just by indexing. Moreover, recovering dense 3D geometry of the eye region from images by fitting parametric models of the shape and texture has been previously proposed [62]. 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 for data augmentation and pseudo-ground truth generation. For instance, [75] used CycleGAN [74] to create a new training corpus in order to balance emotion classes in the task of emotion classification. More recently, GANcraft [23] employed SPADE [50] 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 [15], a face reenactment system. In contrast to person-specific reenactment methods [14,31,36] or person-generic landmark-driven approaches [60,66,67], HeadGAN is able to perform free-view synthesis using a single source image.

3. Methodology

3.1. Problem Definition and Motivation

In this work, our goal is to design a method to predict 3D gaze from images in-the-wild, as a by-product of estimating dense 3D eye meshes. That is, given a face image I, our model estimates \(2 \times N_v\) 3D coordinates \(V = [V_l^T, V_r^T]^T\), where \(V_l \in \mathbb{R}^{N_v \times 3}\) are coordinates corresponding to the left eyeball while \(V_r \in \mathbb{R}^{N_v \times 3}\) to the right. Inspired by recent work in self-supervised 3D body pose estimation [26, 39, 57], we adopt multi-view constraints to train dense 3D eye mesh regression based on face images in-the-wild 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 [15], a recent face reenactment method, to generate novel face poses from existing images. HeadGAN is able to synthesize face animations, using dense face geometry as driving signal and single source images. Using dense geometry guaranties 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 meaningful and consistent eye geometry across different images and datasets, requires establishing a unified 3D representation of eyes. 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, \(M_{l}\) and \(M_{r}\), of the above mesh to represent a left and a right reference eyeball respectively. This representation allows us to allocate semantic labels to different sets of vertices of the eyeball, such as the cornea and iris, as well as retrieving sparse point sets, such as the iris border (Fig. 3 (a)).

3D Eyes Ground-Truth from Gaze Datasets

When gaze labels are available, as for example in gaze estimation datasets, exact supervision can be acquired by automatically fitting the eyeball template on face images, based on sparse landmarks around the iris and the available gaze labels, as also described in Fig. 3 (b). To create such ground truth data for our experiments, we employed the method of [49] to extract sparse iris landmarks from images, but any similar method could have been used.

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. Having recovered 3D face coordinates in image space, we first align our eyeball templates to the 3D face. Then, we lift 2D iris predictions...
Figure 3. (a) Our rigid eyeball mesh template consisted of $N = 481$ vertices and $T = 928$ triangles. (b) Ground truth data generation pipeline, applied on samples of gaze estimation datasets for which gaze ground truth is available. (c) Pseudo-ground-truth data generation pipeline, applied on arbitrary face images without the need for gaze information. Notice how the “corrected eyeballs” (the result of the pipeline) are visually closer to the subjects true gaze than the original “aligned eyeballs”.

to 3D by finding the nearest vertexes from the aligned 3D eye templates (we use only $x$ and $y$ coordinates to find the nearest neighbours). 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 [12] and for 2D iris localization [49] as above. The process is graphically presented in Fig. 3 (c).

### 3.3. 3D Eye Mesh 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 $128 \times 128 \times 3$ and stack them channel wise along with a cropped image of the face. We employ a simple model architecture consisting of a ResNet-18 [25] to extract features, followed by two fully connected layers to map features to two separate eye modalities, which are a) dense 3D eye coordinates and b) a 3D gaze vector. As final gaze output, we consider the mean direction calculated from the two modalities. From our experiments, training with eye coordinate targets alone, usually resulted in similar or slightly worse results than training with gaze direction targets. However, when combining both eye modalities we were able to achieve the best model performance, as also discussed in Sec. 4.2.

To train the above network we employ standard 3D mesh regression losses, as well as gaze constraints to boost gaze accuracy. In particular, for mesh regression we enforce a vertex loss and an edge length loss between the model outputs and the respective ground truth or pseudo-ground truth. The vertex loss can be expressed as follows:

$$L_{\text{vert}} = \frac{1}{N_v} \sum_{j=\{l,r\}} \sum_{i=1}^{N_v} \| V_{j,i} - V_{j,i}^* \|_1,$$  (1)

where $V_{j,i} \in \mathbb{R}^{N_v \times 3}$ and $V_{j,i}^* \in \mathbb{R}^{N_v \times 3}$ 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:

$$L_{\text{edge}} = \frac{1}{3N_t} \sum_{j=\{l,r\}} \sum_{i=1}^{3N_t} \| E_{j,i} - E_{j,i}^* \|_2,$$  (2)

where $E_{j,i} \in \mathbb{R}^{3N_t}$ and $E_{j,i}^* \in \mathbb{R}^{3N_t}$ 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 the following gaze loss to the gaze output of our model:

$$L_{\text{gaze}} = \frac{1}{3} \arccos(g^T g^*)$$  (3)

where $g$ and $g^*$ are the model output and the gaze (pseudo-)ground truth respectively. We combine losses Eq. (1), Eq. (2) and Eq. (3) in a single loss function to train eye 3D mesh regression models with supervision from ground truth or pseudo-ground truth coordinates and gaze vectors. The combined loss is written as:

$$L_{(P)\text{GT}} = \lambda_v L_{\text{vert}} + \lambda_e L_{\text{edge}} + \lambda_g L_{\text{gaze}},$$  (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 $\lambda_v = 0.1$, $\lambda_e = 0.1$ 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 models to generalize to unseen domains, as can be seen by our experiments in Sec. 4.4.1. However,
automatically generated 3D eyes and gaze labels include errors and inconsistencies which are hard to identify and filter out, when no real gaze information is available. To balance 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 pose, generated by HeadGAN as described in Sec. 3.2.

Recovering dense 3D face coordinates and pose from images has recently been quite reliable [1, 12, 19]. Having a pair of images $I_1$ and $I_2$ of the same subject and their reconstructed 3D faces, we can compute a transformation matrix $P \in \mathbb{R}^{3 \times 4}$ 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 additionally supervise 3D regression of eyes without depending on ground truth. That is, we are able to restrict our model’s predictions to be consistent over an image pair, as output vertices should coincide when transformation $P$ is applied to one of the pair’s outputs. Particularly, we form the following pair vertex loss:

$$
L_{MV,\text{vertex}} = \frac{1}{N_v} \sum_{j=\{l,r\}} \sum_{i=1}^{N_v} \|V_{1,j,i}P_p - V_{2,j,i}\|_1, \tag{5}
$$

where $V_{1,j,i}, V_{2,j,i} \in \mathbb{R}^{N_v \times 4}$ for $j = \{l, r\}$ are the output matrices for left and right eyes, which correspond to input images $I_1$ and $I_2$. $V_{1,j,i}, V_{2,j,i} \in \mathbb{R}^4$ 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 $P$ to scale $s$, rotation $R$ and translation $t$ components and employ $R$ in the following loss:

$$
L_{MV,\text{gaze}} = (180/\pi) \arccos((g_1^TR^T)g_2) \tag{6}
$$

where $g_1$ and $g_2$ are the model outputs for input images $I_1$ and $I_2$ respectively. We combine losses Eq. (5), and Eq. (6) in a single loss function to enforce multi-view consistency in mesh regression and direct gaze estimation, between model outputs coming from pairs of input images. The combined loss is written as:

$$
L_{MV} = \lambda_{MV,v}L_{MV,\text{vertex}} + \lambda_{MV,g}L_{MV,\text{gaze}}, \tag{7}
$$

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 three available supervision signals, i.e. ground truth ($L_{GT}$), pseudo-ground truth ($L_{PGT}$) and multi-view supervision ($L_{MV}$), we utilize the following overall loss function:

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

which regularization parameters $\lambda_{GT}$, $\lambda_{PGT}$ and $\lambda_{MV}$ all set to one.

4. Experiments

4.1. Datasets

Gaze Datasets Collected in a lab environment, ETH-XGaze [70] consists of 756K frames of 80 subjects and
includes large head pose and gaze variation. Collected in uncontrolled indoor environments by the frontal camera of mobile devices, MPIIFaceGaze [71] includes smaller head pose and gaze variation and consists of 45K images of 15 subjects, while GazeCapture [37] contains almost 2M frontal face images of 1474 subjects. In contrast to the above datasets, Gaze360 [30] 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 and gaze as well as the variation in lighting and backgrounds, makes Gaze360 the most challenging yet appropriate benchmark for in-the-wild gaze estimation available in literature.  

**In-The-Wild Face Datasets** In contrast to gaze datasets, in-the-wild face datasets consist of significantly more unique subjects and capturing environments. For our experiments, we employed a collection of four publicly-available datasets FFHQ [28] (70K images), AFLW [33] (25K images), AVA [20, 42, 43] and CMU-Panoptic [27]. Among these, 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 a lab environment and captures interactions of multiple people in the same scene.

FFHQ and AFLW include few faces per image and thus, all faces are employed. Instead, AVA and CMU-Panoptic include frames with multiple faces. From these, we randomly select 80K faces from each dataset with maximum yaw angle in head pose of 90°. Similarly to [35], for CMU we employed only frames captured with cameras in eye height. We name this collection of 255K images as “In-The-Wild Gaze” dataset (ITWG). Lastly, to enforce multi-view supervision as described in Sec. 3.4, we synthesized one novel view for each image 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 in our experiments.

### 4.2. Gaze Estimation via 3D Eye Reconstruction

Here we experimentally evaluate our suggestion that gaze estimation accuracy benefits from replacing the training target from gaze vectors or angles to 3D dense 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 four commonly utilized gaze databases, namely MPIIFaceGaze (MPII) [71], GazeCapture (GC) [37], Gaze360 (G360) [30] and ETH-XGaze (EXG) [70] for which specific data split for training and testing are provided. In particular, for Gaze360 we consider only the frontal facing images with yaw angle up to 90°.

We compare against state-of-the-art methods [9, 11, 30, 35, 45, 47, 70, 72] and demonstrate that by combining training for 3D dense coordinates and gaze instead of just gaze vectors or angles, we are able to reach or beat their performance. We report results in Tab. 1. It is worth noting here that our methods employ a simple network architecture and training pipeline, while most methods consist of elaborate models or training schemes, designed to improve gaze accuracy. From the results we observe that when coordinates are employed alone performance slightly drops. However, when both eye modalities are employed performance overall increases. Thus, for the rest of our experiments we employ the combined target training scheme.

Table 1. Comparison between state-of-the-art gaze estimation methods and our models trained with gaze vectors (v), meshes (m) and the combination of the two (m+v), on within-dataset experiments. Combining the two modalities during training leads to better performance in the 3 out of 4 tested datasets.

| Dataset | Compared Methods | Supervision Type |
|---------|------------------|------------------|
| MPII    | 4.04            | 4.00             |
| G360    | 10.7            | 10.6             |
| GC      | -               | 14.9             |
| EXG     | -               | 11.1             |
| ITWG-MV| -               | 3.49             |

\[ \text{Table 1. Comparison between state-of-the-art gaze estimation methods and our models trained with gaze vectors (v), meshes (m) and the combination of the two (m+v), on within-dataset experiments. Combining the two modalities during training leads to better performance in the 3 out of 4 tested datasets.} \]
4.3. Multi-view Supervision Method Evaluation

In this section, we evaluate our multi-view supervision approach in both within-dataset and cross-dataset experiments. We believe that [35] is the most similar method to ours, as it is the only one in literature that attempts to improve gaze estimation generalization by utilizing in-the-wild face data. Additionally, even though GazeOnce [69] proposes an interesting method in this line of research, the presented experiments are conducted on the accompanied dataset MPSGaze which has not become available yet.

Cross-dataset Evaluation We design two cross-dataset experiments to test the performance of our method on Gaze360 (G360) and report the results on Tab. 2(a) and (b). Particularly, the experiments are the following: 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 GazeCapture (GC).

From the results on the above experiments, it becomes obvious that our geometry-aware pseudo-labels employed within our multi-view supervision training effectively support gaze estimation generalization to unseen domains, even without any available ground truth. In particular, in experiment a) our method outperforms [35] by 30% with AVA, 12% with CMU, 9.5% with AVA+CMU and 14% when we employ our large-scale ITWG-MV. Similarly, in experiment b) when additionally employing ground truth from GC our method outperforms [35] by 7%. Overall, our method always outperforms the compared one, while it yields significant improvement when no ground truth is available.

Within-dataset Evaluation Here we compare our method against state-of-the-art in within-dataset gaze estimation on Gaze360 (G360). Similarly to [35], we employ AVA for additional supervision, while we also examine the effect of the larger-scale ITWG-MV. The results, presented in Tab. 2 (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.

| Dataset | [35] | Ours | Dataset | [35] | Ours | Dataset | [35] | Ours |
|---------|------|------|---------|------|------|---------|------|------|
| AVA     | 29.0 | 22.4 | GC      | 30.2 | 29.2 | G360    | 11.1 | 10.1 | 9.6   |
| CMU     | 26.0 | 20.3 | GC+AVA  | -    | 19.5 | G360+AVA| -    | 10.2 | 9.7   |
| CMU+AVA | 22.5 | 19.7 | GC+AVA+CMU | -    | -    | G360+AVA+CMU | -    | -    | 9.5   |
| ITWG-MV | 18.4 | -    | GC+ITWG-MV | -    | 17.6 | G360+ITWG-MV | -    | -    | 9.3   |

Table 3. 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 accuracy is in degrees (lower is better).

| Dataset | Pseudo-GT | Multi-View loss | G360 | GC | EXG | MPII |
|---------|-----------|-----------------|------|----|-----|------|
| ITWG    | ✓         | -               | 23.1 | 14.8 | 24.3 | 13.6 |
| ITWG-MV | -         | ✓               | 47.4 | 33.2 | 41.1 | 32.8 |
| 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  | 21.5 | 8.6  |

4.4. Ablation studies

4.4.1 Gaze Pseudo-Labels and Multi-View Supervision

In this Section, 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, b) training with ITWG-MV utilizing only the multi-view consistency constraints and no pseudo-labels and c) training with ITWG-MV while employing both pseudo-labels and the multi-view consistency loss. To further evaluate the effect of the pseudo-labels and multi-view loss, we repeat the above experiments by adding ground truth supervision from GazeCapture (GC). We test our models on the frontal facing images of the test set of Gaze360 (up to 90° yaw), as well as in GazeCapture (GC), ETH-XGaze (EXG) and MPIIFaceGaze (MP), and report the results in Tab. 3.
Table 4. The effect of head pose variation of ITWG. Incorporating data with more diverse head pose, leads to increasingly higher accuracy (measured in degrees, lower is better.)

| Training Dataset | ITWG Yaw Variation |
|------------------|---------------------|
| MPII             | 26.7                |
| MPII+ITWG-MV     | 25.6 21.9 19.6 18.4 |

ITWG-MV leads to very high errors which is reasonable as no supervision for the eyeball topology exists.

4.4.2 The Effect of Head Pose Distribution of ITWG

Head pose variation differences between the train and test set, is one of the main reasons that gaze-estimation models fail in cross-dataset validation. 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. Here, we study the effect of the head pose variation of ITWG, by employing subsets of ITWG with various levels of head pose variation and conducting cross-dataset experiments. In particular, we consider four subsets of ITWG, with maximum yaw angle of $5^\circ$, $20^\circ$, $40^\circ$ and $90^\circ$ (all) respectively.

We train the full version of our model with ground truth supervision from MPIIFaceGaze (MPII) as well as pseudo-labels and multi-view supervision from the four versions of ITWG-MV. The model is tested on the frontal facing images of the test set of Gaze360 (up to $90^\circ$ yaw) and results are reported in Tab. 4, from which it becomes obvious that the more diverse ITWG is with respect to head pose, the lower the gaze estimation error becomes. This effect can be graphically demonstrated by plotting the gaze predictions of the different models over the ground truth values of the test set. Starting with very limited prediction variation in Fig. 6 when only MPII is used for training, ground truth is represented increasingly better when more diverse versions of ITWG are utilized. Further analysis is included in the supplemental material.

4.5. Implementation Details

Data Augmentation. As mentioned above we center crop 3 patches (left eye, right eye and face) resize them to shape $128 \times 128 \times 3$ and stack them channel-wise. Before stacking the image patches we randomly scale, translate, flip, and add noise on the color channel with probability 0.5. Due to different image quality on each dataset we also add Gaussian blur to the image. The intensity of the blur is randomly selected from a set of kernels. Note that we augment each patch with the same augmentation parameters.

Training Details. We train our method using a Adam optimizer (weight decay at 0.0005, and batch size of 128) on a single Tesla V100-PCIE (32GB) GPU. The learning rate starts from 1e-6, linearly warming up to 1e-4 in the first 3 epochs and then divided by 10 at 60 and 80 epochs. The training process is terminated at 100 epochs.

5. Limitations

In the experiments of Sec. 4, we have shown that pseudo-ground truth can be effectively utilized in gaze estimation. However, our gaze pseudo-labels do not come without any limitations. From our experimentation, we have identified that our pseudo-label generation pipeline results in weaker labels in the pitch axis. Another drawback of this pipeline is that the outputs are highly related to the accuracy of 3D face and 2D iris alignment. Lastly, one limitation of our overall method is that it cannot operate on images without a visible face (when the face is looking away from the camera), which constitutes an area for future extensions.

6. Conclusion

In this work we present a novel, weakly-supervised method for gaze estimation, 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, results in better performance. 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 supervision framework for effective training. Employing both, we achieve improvements in cross-dataset and within-dataset experiments.
References

[1] Vítor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In CVPR, 2021. 5

[2] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In HRI, 2014. 1

[3] Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In CVPR, 2022. 2

[4] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In CHI, 2020. 1

[5] Yujun Cai, Lihao Ge, Jianfei Cai, and Jiandong Niu. Weakly-supervised 3d hand pose estimation from monocular rgb images. In ECCV, 2018. 2

[6] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Judith Richter, Thérèse Eder, Fabian Hüttig, Constanze Keutel, and Enkelejda Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In ACM ETRA, 2020. 1

[7] Meixu Chen, Yize Jin, Todd Goodall, Xiangyu Yu, and Alan Conrad Bovik. Study of 3d virtual reality picture quality. IEEE Journal of Selected Topics in Signal Processing, 2020. 1

[8] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In AAAI, 2020. 2

[9] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In ICPR, 2022. 6

[10] Yihua Cheng, Feng Lu, and Xiong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In ECCV, 2018. 2

[11] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. arXiv preprint arXiv:2104.12668, 2021. 6

[12] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In CVPR, 2020. 1, 4, 5, 6

[13] Yu Deng, Jiaolong Yang, Sichen 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. 2

[14] Michail Christos Doukas, Mohammad Rami Koujan, Viktoria Sharmskaya, Anastasios Roussos, and Stefanos Zafeiriou. Head2head++: Deep facial attributes re-targeting. T-BIOM, 2021. 3

[15] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoria Sharmskanska. Headgan: One-shot neural head synthesis and editing. In ICCV, 2021. 2, 3

[16] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Retgane: Real-time eye gaze estimation in natural environments. In ECCV, 2018. 2

[17] Wolfgang Fuhl, Gjergji Kasneci, and Enkelejda Kasneci. 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. 2

[18] 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, 2014. 2

[19] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In CVPR, 2019. 5

[20] Chunhui Gu, Chen Sun, David A. Ross, Carl Von-der-Brck, 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. 6

[21] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In CVPR, 2019. 1

[22] Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. Domain adaptation gaze estimation by embedding with prediction consistency. In ACCV, 2020. 2

[23] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. GANcraft: Unsupervised 3D Neural Rendering of Minecraft Worlds. In ICCV, 2021. 3

[24] Junfeng He, Khoi Pham, Nachiappan Valliappan, Pingmei Xu, Chase Roberts, Dmitry Lagun, and Vidhya Navalpakam. On-device few-shot personalization for real-time gaze estimation. In ICCV Workshops, 2019. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 4

[26] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In CVPR, 2020. 2, 3

[27] 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, 2015. 6

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In CVPR, 2019. 6

[29] Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In ECCV, 2022. 1

[30] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In ICCV, 2019. 2, 6, 7

[31] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. TOG, 2018. 3

[32] Chris L. Kleineke. Gaze and eye contact: a research review. Psychological bulletin, 1986. 1

[33] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-
scale, real-world database for facial landmark localization. In ICCVW, 2011.

[34] Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. In TOG, 2019.

[35] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. In CVPR, 2021.

[36] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Rousséos, and Stefanos Zafeiriou. Head2head: Video-based neural head synthesis. In FG, 2020.

[37] Kyle Krafft, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In CVPR, 2016.

[38] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In CVPR, 2020.

[39] Yang Li, Kan Li, Shuai Jiang, Ziyue Zhang, Congzhen-tao Huang, and Richard Yi Da Xu. Geometry-driven self-supervised method for 3D human pose estimation. In AAAI, 2020.

[40] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation with calibration. In BMVC, 2018.

[41] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In ICCV, 2021.

[42] Manuel J. Marín-Jiménez, Vicky Kalogeiton, Pablo Medina-Suárez, and Andrew Zisserman. LAEO-Net++: Revisiting people looking at each other in videos. TPAMI, 2021.

[43] Manuel J. Marín-Jiménez, Vicky Kalogeiton, Pablo Medina-Suárez, and Andrew Zisserman. Laeo-net: Revisiting people looking at each other in videos. In CVPR, 2019.

[44] Aman George Mavely, J. E. Judith, P. A. Sahal, and Steffy Ann Karuvilla. Eye gaze tracking based driver monitoring system. In ICCS, 2017.

[45] Jun O Oh, Hyung Jin Chang, and Sang-II Choi. Self-attention with convolution and deconvolution for efficient eye gaze estimation from a full face image. In CVPRW, 2022.

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

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

[48] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In ECCV, 2018.

[49] 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, 2018.

[50] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In CVPR, 2019.

[51] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando de la Torre, and Yaser Sheikh. Audio- and gaze-driven facial animation of codec avatars. In WACV, 2021.

[52] Iasonas Kokkinos Riza Alp Guler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. In CVPR, 2018.

[53] 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, 2013.

[54] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3D gaze estimation. In CVPR, 2014.

[55] Yunjia Sun, Jiaiwei Zeng, Shiguang Shan, and Xilin Chen. Cross-encoder for unsupervised gaze representation learning. In ICCV, 2021.

[56] Melodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. Wearable eye tracking for mental health monitoring. Computer Communications, 2012.

[57] Bastian Wandt, Marco Rudolph, Pettrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3D human pose estimation in the wild. In CVPR, 2021.

[58] Kang Wang, Rui Zhao, and Qiang Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In CVPR, 2018.

[59] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In CVPR, 2019.

[60] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In NeurIPS, 2019.

[61] Yaoming Wang, Yangzhou Jiang, Lin Li, Bingbing Ni, Wenui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In CVPR, 2022.

[62] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. A 3D morphable eye region model for gaze estimation. In ECCV, 2016.

[63] Yu Yu, Gang Liu, and Jean-Marc Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In ECCV Workshops, 2018.

[64] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In CVPR, 2019.

[65] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In CVPR, 2020.

[66] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lemptisky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In ECCV, 2020.

[67] E. Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Stefanos Zafeiriou. Head2head: Video-based realistic head avatars. In ECCV, 2020.

[68] Jichao Zhang, Jingjing Chen, Hao Tang, Wei Wang, Yan Yan, Enver Sargineto, and Nicu Sebe. Dual in-painting model for unsupervised gaze correction and animation in the wild. In ACM MM, 2020.
[69] Mingfang Zhang, Yunfei Liu, and Feng Lu. Gazeonce: Real-time multi-person gaze estimation. In CVPR, 2022. 3, 7
[70] 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. 2, 5, 6, 7
[71] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In CVPR, 2015. 2, 6
[72] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It’s written all over your face: Full-face appearance-based gaze estimation. In CVPRW, 2017. 6
[73] 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, 2017. 2
[74] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In ICCV, 2017. 3
[75] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. Emotion classification with data augmentation using generative adversarial networks. In PAKDD, 2018. 3
**1. Model Demo**

To showcase the effectiveness of our method in real conditions, we have trained a general gaze estimation model based on 3D eye mesh regression. The model has been trained with the four gaze datasets discussed in this paper (MPIIFaceGaze [7], Gaze360 [3], ETH-XGaze [6] and GazeCapture [4]), as well as the in-the-wild images with pseudo-labels from our ITWG-MV dataset which we employ within our proposed framework. Combining the accuracy of gaze datasets and the diversity of ITWG-MV, our model is robust to a wide variety of scenarios, e.g. challenging head poses, lighting settings and occlusions, that are commonly encountered in real scenarios. A live demo of our method can be found here [http://35.92.148.85:7011/](http://35.92.148.85:7011/).

**2. Additional Ablation on the effect of ITWG Head Pose Variation**

In this section we present further analysis on the effect of head pose variation of ITWG on the model’s generalization ability (Sec. 4.4.2 of the main paper). For training we employ different subsets of ITWG based on head pose (< 5°, < 20°, < 40°, all) and additional ground truth supervision from MPIIFaceGaze (MPII) or GazeCapture (GC). We present results of our models for different subsets of Gaze360’s test set, based again on the head pose yaw angle (< 5°, < 20°, < 40°, 90°).

Results reported in Tab. 1 demonstrate that an improvement of 8° to 13° (31% to 37%) is achieved in all cases (for all subsets of Gaze360) between the baselines of training just with gaze datasets (MPII, GC) and our full method of including ITWG and multi-view supervision in training. Another interesting fact is that performance consistently increases when training with more diverse subsets of ITWG in all cases. Lastly, it is worth noticing that even though performance increase is expected for the subsets of Gaze360 with large head pose values (> 40°), as MPII and GC do not include such images at all, the larger increase in performance is seen for near-frontal images. This fact, validates the quality of our pseudo-labels in near-frontal head poses and the effectiveness of the overall pseudo-labelling method and our multi-view supervision algorithm.

**3. Additional Ablation on the effect of Pseudo-Labels and Multi-View Supervision**

In this section we present further evaluations on the effect of our pseudo-labels and multi-view consistency constraints during training. To this end, we repeat the experiments described in Sec. 4.4.1 of the main paper for 3 additional cases of available ground truth supervision and report the results in Tab. 2. In particular, we utilize Gaze360 (G360), ETH-XGaze (EXG) and MPIIFaceGaze (MPII) as source datasets with valid ground truth. From the results we can draw the conclusion that our method is always effective when there are large differences between the source and target dataset head pose and gaze variation. In such cases ITWG helps to close the gap and results in significant improvement. Only small improvement is noticed for the within-dataset experiment on Gaze360, while results do not improve for within-dataset experiments on ETH-XGaze and MPII.

**4. Calculating Gaze Direction from 3D Eye Meshes**

In this work we have proposed a method to estimate 3D eye meshes from images and employ them for gaze esti-
### 5. Qualitative Results for 3D Gaze Estimation

Here we visualize gaze predictions of our model for training scenarios discussed in Sec. 4.4.2 of the main paper. In particular, we present the results of our model in the two edge cases of Tab. 4 of the main paper, i.e. a) only MPIIFaceGaze is employed for training and b) MPIIFaceGaze and the whole ITWG are employed for training within our multi-view supervision framework. Testing is performed on the images of Gaze360. Fig. 2 includes results for the two cases as well as the ground truth labels of Gaze360. As can be seen, the predicted gaze directions are much closer to the real ones when in-the-wild face data from ITWG are employed for supervision. Especially for profile views, the effect of the pseudo-labels is visually significant.

To further evaluate our method, we apply the above two models on arbitrary in-the-wild face images and present the results on Fig. 3. As actual gaze accuracy cannot be measured for such images (ground truth data are not available), we attempt to draw conclusions based on observation. From the visualizations it can be seen that for side and profile views, our multi-view supervision method (MPII + ITWG) performs substantially better, while for near-frontal ones the predictions improve when occlusions are present (glasses, hair). These qualitative results demonstrate that our method improves cross-domain gaze generalization, without using any additional gaze information or data from the target domains.

### 6. Discussion

In this section we discuss limitations of our work, as well as possible solutions. One common challenge in dense 3D mesh regression is handling occlusions. In our case the biggest challenge comes from faces in profile pose and cases of people wearing glasses. Some examples of applying our model on such cases are depicted in Fig. 4. While transparent glasses do not pose a significant challenge, profile faces make one of the two eyes invisible and sunglasses cover the eyes completely, which causes our model’s accuracy to drop. These are challenges that our method attempts to solve by incorporating in-the-wild data in training. We believe that if progress continues to be made towards this direction, gaze estimation models will gradually become more accurate and robust for in-the-wild scenarios, using less annotated data.

Another limitation of our method lies in the use of synthetic images for weakly-supervised training. For this algorithm, we assume that images of the same subject with different pose but the same difference between head and eye orientation are available. To acquire such data we employ a face reenactment method, HeadGAN [2], which animates the human head given single input images. However, relying on synthetic data for training means that performance is compromised by the quality of image generation. Higher quality of face image synthesis, could lead to better performance for our method.

Lastly, another limitation of our model is that it does not consider the anatomical differences of eyes between people. In more detail, an offset angle exists between the optical and visual axes of eyes according to their real anatomy as shown in Fig. 5. This angle is subject-dependent and usually mentioned as the kappa coefficient of the eyes. Some methods have attempted to model this offset or incorporate it in their...
Figure 2. Results of our models trained with MPIIFaceGaze (blue vectors) and combined MPIIFaceGaze and ITWG with multi-view supervision (yellow vectors), applied on the test set of Gaze360 (red vectors). The predicted gaze directions are closer to the ground truth when ITWG is included in training. Especially for side and profile views, the effect of the pseudo-labels is visually significant.

Figure 3. Results of our models trained with MPIIFaceGaze (blue vectors) and combined MPIIFaceGaze and ITWG with multi-view supervision (yellow vectors), applied on arbitrary in-the-wild images. Our full model predicts robust gaze labels across all head pose angles for yaw in $[-90^\circ, 90^\circ]$. Especially for side and profile views, as well as for faces with occlusions (glasses, hair, sunglasses), the effect of the pseudo-labels is visually significant.

In our method, 3D gaze predictions are calculated by the orientation of the central axis of our 3D eyeball template, which coincides with the optical axis of the human eyes. To make our system robust...
Figure 4. Results from applying our model on faces with glasses (1st row), sunglasses (2nd row) and faces in near-profile pose (3rd row). While normal glasses and profile faces do not pose a significant challenge, sunglasses are more difficult to handle.

Figure 5. Eyeball anatomy demonstrating the offset between the optical and visual axis.

to variations of face identity, we rely on large in-the-wild face datasets. However, employing an anatomically aware 3D eyeball template or designing a strategy for personalizing our model constitutes an interesting direction for further research.

References

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

[2] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In ICCV, 2021.

[3] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In ICCV, 2019.

[4] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In CVPR, 2016.

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

[6] Xuccong 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.

[7] Xuccong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It’s written all over your face: Full-face appearance-based gaze estimation. In CVPR Workshops, 2017.