

Learning-by-Novel-View-Synthesis for Full-Face Appearance-based 3D Gaze Estimation

Jiawei Qin, Takuru Shimoyama, Yusuke Sugano
 Institute of Industrial Science, The University of Tokyo
 {jqin, tshimo, sugano}@iis.u-tokyo.ac.jp

Abstract

Despite recent advances in appearance-based gaze estimation techniques, the need for training data that covers the target head pose and gaze distribution remains a crucial challenge for practical deployment. This work examines a novel approach for synthesizing gaze estimation training data based on monocular 3D face reconstruction. Unlike prior works using multi-view reconstruction, photo-realistic CG models, or generative neural networks, our approach can manipulate and extend the head pose range of existing training data without any additional requirements. We introduce a projective matching procedure to align the reconstructed 3D facial mesh to the camera coordinate system and synthesize face images with accurate gaze labels. We also propose a mask-guided gaze estimation model and data augmentation strategies to further improve the estimation accuracy by taking advantage of the synthetic training data. Experiments using multiple public datasets show that our approach can significantly improve the estimation performance on challenging cross-dataset settings with non-overlapping gaze distributions.

1. Introduction

Gaze estimation has been considered an important research topic in the computer vision community with many applications. Vision-based techniques have the potential to bring gaze estimation capability to arbitrary cameras. However, despite recent advances in machine learning-based approaches [34, 43, 58, 61, 62], it is still challenging to accurately predict gaze directions under extreme head poses and diverse lighting conditions.

One of the fundamental difficulties is the requirement for an appropriate training dataset. There have been many efforts on creating diverse in-the-wild gaze datasets [12, 26, 27, 61]. However, it is not a trivial task to construct a dataset covering all crucial factors including head pose and gaze distribution, illumination environment, background appear-

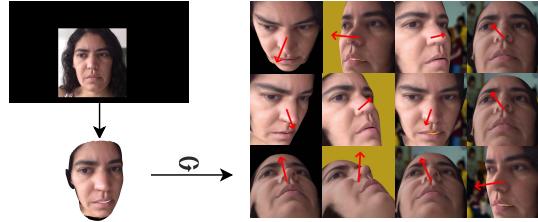


Figure 1. We propose a learning-by-synthesis appearance-based gaze estimation approach based on single-image 3D face reconstruction. Our projective matching procedure aligns the reconstructed face with the ground-truth gaze position for generating precise training data.

ance, demographic diversity, imaging properties, and accurate gaze labels.

As actively studied in other computer vision tasks [21, 38, 39, 42, 46, 51], one potential approach to obtain targeted training data is to use synthetic images.

In the context of appearance-based gaze estimation, accurate ground-truth 3D gaze direction is required when synthesizing images. Previous approaches use either multi-view 3D reconstructed data [15, 43] or hand-crafted eyeball models [50] to synthesize eye images for appearance-based gaze estimation. However, it is still challenging to capture 3D reconstruction data under diverse illumination conditions. While hand-crafted computer graphics models can potentially address this limitation, there is also a vast domain gap from real images [40]. These limitations become more prominent for full-face gaze estimation approach [27, 58, 62]. Generative neural rendering models are one of the promising approaches to generate full-face images while controlling the gaze directions [64], but it is not easy to ensure their labels are accurate enough to be used as training data.

This work proposes an alternative approach to learning-by-synthesis full-face appearance-based gaze estimation via single-image 3D face reconstruction. As illustrated in Fig. 1, we reconstruct 3D facial shapes from existing gaze

datasets and synthesize novel views by rotating the reconstructed faces. However, since most of the single-image 3D face reconstruction methods do not provide physical 3D shapes in the camera coordinate system, it brings another challenge of preserving accurate gaze labels under novel views. To address this issue, we introduce a projective matching procedure to ensure that the reconstructed 3D facial surface is associated with the original camera coordinate system and the ground-truth gaze target position.

In addition, we propose a novel mask-guided gaze estimation model that utilizes face region masks obtained through the rendering process as additional supervision, aiming to take full advantage of the synthetic data. We also propose to render training images with lighting and background augmentation to enhance the diversity of image appearance. We evaluate the performance of the proposed approach on the dataset extrapolation task to cover unseen head poses and gaze directions. By combining our synthetic data and mask-guided estimation model, we show that our approach can outperform the gaze estimation results of other state-of-the-art synthetic- and real-image training datasets.

The contributions of this work are threefold. (i) We propose a novel approach for creating training data for appearance-based gaze estimation via 3D face reconstruction. To our knowledge, this is the first work to prove that single-image face reconstruction outputs can be used to train full-face appearance-based gaze estimation models. (ii) We propose a novel mask-guided soft-attention model for gaze estimation. Together with the data augmentation, our gaze estimation method fully utilizes the nature of synthetic training. (iii) Through experiments, we verify that our approach can successfully extend the gaze range of the source dataset. Our synthetic training dataset provides better model performance than baseline datasets using real and synthetic images.

2. Related Work

Traditional model-based gaze estimation methods use 3D eyeball models with geometric features to infer 3D gaze directions [17, 22]. On the other hand, appearance-based gaze estimation methods directly map the image to gaze direction [45]. Methods in this category have fewer restrictions in the hardware setup and are more suitable for in-the-wild settings.

Most appearance-based methods take eye-only images as input [43, 45, 48, 53, 55, 56, 61], and there have been also some attempts to explore two-eye combination inputs [6, 8, 20, 33]. In contrast, some prior work demonstrated that full-face input can improve the robustness and accuracy of appearance-based gaze estimation [5, 27, 60, 62]. While this work also focuses on the full-face appearance-based gaze estimation task, we explore the potential of us-

ing single-image 3D face reconstruction for synthesizing full-face training data for the first time.

Gaze Estimation Datasets. While some datasets were collected using mobile devices during in-the-wild daily-life situations under diverse illumination conditions, they often suffer from limited head and gaze ranges [24, 27, 61–63]. Some datasets reached higher variety in head pose and gaze by using more complex recording setups, but the environment and illumination are always limited to the controlled conditions [12, 14, 41].

Recent datasets have been collected with further extended diversity in head pose ranges and environment conditions [26, 58]. However, a significant effort is still required to acquire training datasets meeting the requirement of head pose and appearance variations under the deployment environment. This work aims at addressing this issue by providing a method for extending the head pose ranges of source datasets as well as augmenting the environment diversity. **Learning-by-Synthesis for Gaze Estimation.** To address the limitations of real-world data collection, there have been some efforts on creating synthetic training data for appearance-based gaze estimation using multi-view stereo reconstruction [43] or hand-crafted photo-realistic computer graphics models [49, 52]. However, the multiview setup has a fundamental limitation that the environment is fixed to the lab condition, and the domain gap between real and purely synthetic images is not negligible. Zheng *et al.* proposed a neural network for redirecting gaze and head pose, which can be also used to generate synthetic training data [64]. However, such neural rendering models cannot guarantee the facial appearance exactly matches the target gaze label. In this work, we take yet another approach based on single-image 3D face reconstruction for accurate data synthesis.

Domain Adaptation for Gaze Estimation. When using synthetic data, the domain gap between synthetic and real images can be a critical issue. However, in the context of appearance-based gaze estimation, there have been few studies dealing with such an unsupervised, cross-environment domain adaptation task. Fundamentally speaking, there have been few research examples of domain adaptation for regression tasks, only with a few examples designed for head pose estimation [28] and body pose estimation [44]. Shrivastava *et al.* [40] proposed SimGAN, an unsupervised domain adaptation approach that refines synthetic eye images to be visually similar to real images. However, their method was designed for eye images and its effectiveness has never been validated on full-face gaze estimation. Liu *et al.* [30] recently proposed an unsupervised domain adaptation framework based on collaborative learning. Although their work addresses the full-face gaze estimation task, its effectiveness on synthetic source data has not been evaluated. In contrast to these methods taking do-

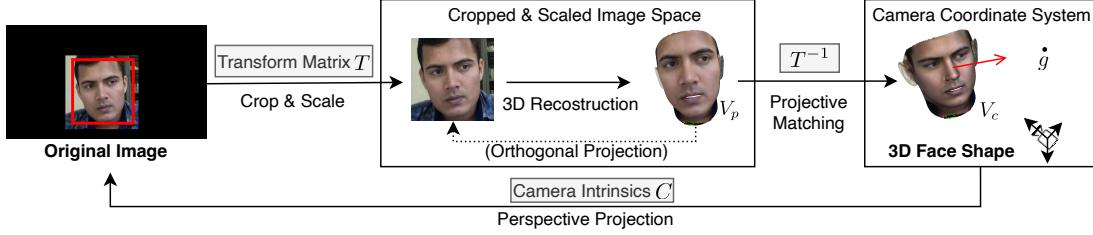


Figure 2. Overview of our data synthesis pipeline. We assume that 3D face reconstruction methods generate facial meshes under an orthogonal projection model, and we convert the mesh via the proposed projective matching to align with the ground-truth gaze position in the input camera coordinate system.

main adaptation approaches, we propose a method that addresses the domain gap by fully utilizing the characteristics of the synthetic training data.

3D Face Reconstruction. Monocular 3D face reconstruction techniques have also made significant progress in recent years [69]. While reconstructed 3D faces have also been used to augment data for face recognition [32, 57, 66], no prior work explored its usage in full-face appearance-based gaze estimation. Methods based on 3D morphable models [9, 47] usually approximate the facial textures via the appearance basis [3, 29, 35, 36], and therefore eye region appearances can be distorted. To preserve accurate gaze labels after reconstruction, this work utilizes 3D face reconstruction methods that sample texture directly from the input image [2, 4, 10, 18, 19, 67, 68]. In addition, since many prior works rely on orthogonal or weak perspective projection models, we discuss how to precisely align the reconstruction results with the source camera coordinate system.

3. View Synthesis via 3D Face Reconstruction

Given an ordinary single-view gaze dataset and 3D face reconstruction results, our goal is to synthesize face images under unseen head poses while preserving accurate gaze direction annotations.

3.1. Overview

Fig. 2 shows the overview of our data synthesis pipeline. We assume that the source gaze dataset consists of 1) face images, 2) the projection matrix (intrinsic parameters) C of the camera, and 3) 3D gaze target position $g \in \mathbb{R}^3$ in the camera coordinate system. Most of the existing gaze datasets contain 3D gaze position annotations [27, 58, 61], and please note that 2-DoF (yaw and pitch) gaze direction representation can also be converted to a 3D position representation by setting a distance to the dummy target.

State-of-the-art learning-based 3D face reconstruction methods usually take a cropped face patch as input and output a 3D facial mesh, which are associated with the input image in an orthogonal or weak perspective projection way. Without loss of generality, we assume that the

face reconstruction method takes a face bounding box defined with center (c_x, c_y) , width w_b , and height h_b in pixels and then resized to a fixed input size by factor (s_x, s_y) . The reconstructed facial mesh is defined as a group of N vertices $V_p = \{\mathbf{v}_p^{(i)}\}_{i=0}^N$. Each vertex is represented as $\mathbf{v}_p^{(i)} = [u^{(i)}, v^{(i)}, d^{(i)}]^\top$ in the right-handed coordinate system, where u and v directly correspond to the pixel locations in the input face patch and d is the distance to the u - v plane in the same pixel unit. Many recent works use this representation [4, 11, 18, 25, 67], and we can convert arbitrary 3D representation to this way by projecting the reconstructed 3D face to the input face patch.

Our goal is to convert the vertices of the reconstructed 3D face V_p to another 3D representation $V_c = \{\mathbf{v}_c^{(i)}\}_{i=0}^N$ where each vertex $\mathbf{v}_c^{(i)} = [x^{(i)}, y^{(i)}, z^{(i)}]^\top$ is in the original camera coordinate system so that it can be associated with the gaze annotation g . In this way, the gaze target location can be also represented in the facial mesh coordinate system, and we can render the facial mesh under arbitrary head or camera poses together with the ground-truth gaze direction information.

3.2. Projective Matching

Since u and v of the reconstructed vertices \mathbf{v}_p are assumed to be aligned with the face patch coordinate system, \mathbf{v}_c has to be on the back-projected ray as

$$\mathbf{v}_c = \lambda \frac{\mathbf{C}^{-1} \mathbf{p}_o}{\|\mathbf{C}^{-1} \mathbf{p}_o\|} = \lambda \frac{\mathbf{C}^{-1} \mathbf{T}^{-1} \mathbf{p}}{\|\mathbf{C}^{-1} \mathbf{T}^{-1} \mathbf{p}\|}, \quad (1)$$

where $\mathbf{p}_o = [u_o, v_o, 1]^\top$ and $\mathbf{p} = [u, v, 1]^\top$ indicates the pixel locations in the original image and face patch in the homogeneous coordinate system, respectively, and

$$\mathbf{T} = \begin{bmatrix} s_x & 0 & -s_x(c_x - \frac{w}{2}) \\ 0 & s_y & -s_y(c_y - \frac{h}{2}) \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

represents the cropping and resizing operation to create the face patch, i.e., $\mathbf{p} = \mathbf{T} \mathbf{p}_o$. λ indicates scaling along the

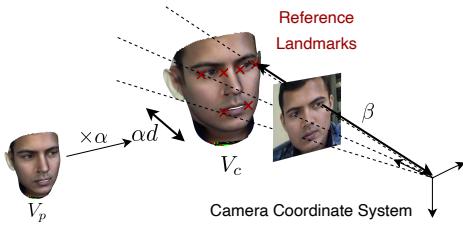


Figure 3. Determining the location of V_c via parameters α and β . α indicates a scaling factor from the pixel to physical (e.g., millimeter) unit, and β is the bias term to align αd to the camera coordinate system.

back-projection ray, and physically means the distance between the camera origin and v_c .

Since Eq. (1) does not explain anything about d , our task can be understood as finding λ which also maintains the relationship between u , v , and d . Therefore, as illustrated in Fig. 3, we propose to define λ as a function of d as $\lambda = \alpha d + \beta$. α indicates a scaling factor from the pixel to physical (e.g., millimeter) unit, and β is the bias term to align αd with the camera coordinate system. Please note that α and β are constant parameters determined for each input image and applied to all of the vertices from the same image.

We first fix α based on the physical size of human faces using a reference 3D face model. More specifically, we use the distance between two eye centers (midpoints of two eye corner landmarks) to determine the scale between pixel and physical units. 3D face reconstruction methods usually require facial landmark detection as a preprocessing, and we can naturally assume that we know the corresponding vertices in V_p to the eye corner landmarks. We use a 68-landmarks 3D face model (taken from the OpenFace library [1]) as our reference. We set $\alpha = l_r/l_p$, where l_p is the distance between eye center vertices in V_p and l_r indicates the eye center distance computed from the reference model.

We then determine β by aligning the reference landmark depth in the camera coordinate system. In this work, we use the face center as the reference, which is defined as the centroid of eyes and mouth corner landmarks. This definition follows a prior work on full-face gaze estimation [59, 62], and we use the same face center as the origin of the gaze vector through the data normalization process and gaze estimation task.

We estimate β as the distance between ground-truth 3D reference location and the scaled reconstructed location as $\beta = \|\bar{v}\| - \alpha \bar{d}$. \bar{d} is the reconstructed depth values computed as the mean of six vertices corresponding to the eyes and mouth corner landmarks obtained in the similar way as when computing α . \bar{v} is the centroid of the 3D locations of the same six landmarks in the camera coordinate system, which are obtained by minimizing the projection error of

the reference 3D model to the 2D landmark locations using the PnP algorithm [13].

3.3. Training Data Synthesis

Once we obtain the facial mesh V_c in the original camera coordinate system, it becomes possible to render the facial mesh under arbitrary head poses together with the ground-truth gaze vector. If our goal is to render a face image in a new camera coordinate system which is defined with extrinsic parameters R_e, t_e , vertex v_c and gaze target position g are both projected to the new coordinate system as $R_e v_c + t_e$ and $R_e g + t_e$ in the same manner. Similarly, if the goal is to render a face image with a target head pose¹ R_t, t_t in a new camera coordinate system given the source head pose R_s, t_s , we can transform the vertices and gaze position as $R_t(R_s)^{-1}(v_c - t_s) + t_t$.

In this work, we further augment the images in terms of lighting conditions and background appearances by virtue of the flexibility in rendering process. To augment the background appearance, we set the background to be random color or random scene images.

Although most of the 3D face reconstruction methods do not reconstruct lighting and albedo, we also maximize the diversity of rendered images by controlling the global illumination. We randomly reduce the ambient light intensity to render darker weak-light images. Fig. 4 shows some examples of the synthesized images using MPIIFaceGaze [62] and ETH-XGaze datasets [58].

3.4. Rendering Details

In the experiments, we applied 3DDFA [18] to reconstruct 3D faces from the source dataset. After projective matching, we rendered new images using the PyTorch3D library [37]. We set the background to be a random RGB value or scene image by modifying the blending setting. In the PyTorch3D renderer, the ambient color $[r, g, b]$ represents the ambient light intensity, ranging from 0 to 1, in which 1 is the default value for full-lighting. For weak-light images, we set them to be a random value between 0.5 and 1. Overall, among all generated images, the ratio of black, random color, and random scene are set to 1:1:3, and half of them are weak lighting. The random scene images are taken from the Places365 dataset [65] and we apply blurring to them before rendering faces.

4. Mask-Guided Gaze Estimation

While the data synthesis process described above can render realistic face regions with accurate gaze direction labels, there still remains a huge domain gap between synthetic and real images. We cannot fully ignore the influence

¹Head pose is defined as the rotation and translation from the face coordinate system to the camera coordinate system.

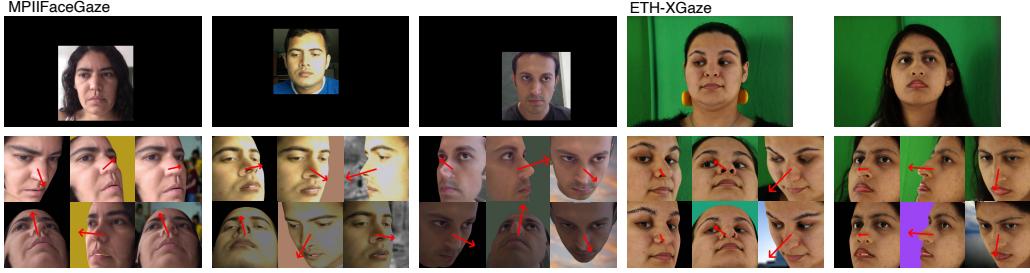


Figure 4. Examples of the synthesized images. The first row shows the source images from MPIIFaceGaze [62] and ETH-XGaze [58]. The second and third rows show synthesized images for MPIIFaceGaze, and pairs of real and synthesized images for ETH-XGaze. For MPIIFaceGaze, the second and third rows show synthesized images in full-light and weak-light. For ETH-XGaze, the second row shows the real images from the dataset, and the third row shows our synthetic images with the same head poses as the second row. For each synthetic example, the three columns show the black, color, and scene image background in turn, and the red arrows indicate gaze direction vectors.

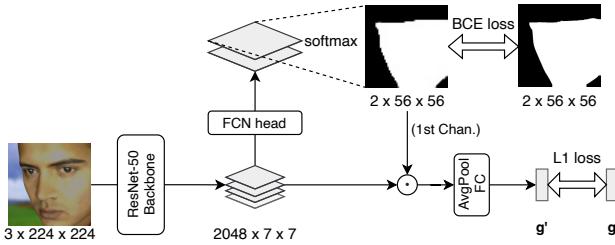


Figure 5. Architecture of the proposed mask-guided gaze estimation network. In addition to the base gaze estimation network, it has an extra segmentation branch whose output segmentation mask serves as a soft attention.

of background and non-face (*e.g.*, hair and clothes) regions in the full-face estimation task. It is also difficult to synthesize invisible face regions in the original image even with the state-of-the-art face reconstruction methods. In this section, we describe our mask-guided gaze estimation model to address the domain gap issue using additional supervision obtained from our synthetic data.

4.1. Network Architecture

Fig. 5 shows the overview of the proposed mask-guided gaze estimation network. When synthesizing the training data, we propose to generate binary masks representing the reliable regions of the reconstructed facial mesh, *e.g.*, frontal face regions which are visible in the source image. In addition to the base gaze estimation network, our proposed architecture has an extra fully-convolutional branch [31] after the feature extraction block to predict segmentation masks corresponding to such synthetic binary masks. The output segmentation mask is then applied to the feature map, serving as a soft attention [54] to enhance informative feature regions.

The network is trained in a multitask manner by combin-

ing two loss functions as $L = L_{\text{gaze}} + \gamma L_{\text{mask}}$, where L_{gaze} and L_{mask} correspond to loss terms evaluating gaze direction and segmentation mask, respectively. Following [58], L_{gaze} is defined as a $L1$ loss between ground-truth and predicted gaze directions. L_{mask} is defined as a binary cross entropy loss between the ground-truth region mask and the predicted segmentation mask.

4.2. Implementation Details

Together with the synthetic face images, mask images are also generated to supervise our mask-guided network. Since most of the faces in the reconstruction source images are nearly frontal, we explicitly use the frontal face region mask for this purpose. We define the frontal face region based on the 2D landmark locations on the input image. Since the reconstructed 3D surfaces are aligned with the input image, we use the 2D landmark locations to define the outline of the face region and filter out the 3D vertices outside the region. We also filter out the vertices with depth values larger than that of the landmark near the jaw and finally obtain the face-region-only vertices for rendering binary masks.

In the following experiments, we use the ImageNet pre-trained ResNet-50 [23] as the backbone network. The output is fed into a 1×1 convolution layer followed by a transposed convolution layer with 16×16 kernel, stride of 8, and padding of 4. After a softmax layer, we obtain the $2 \times 56 \times 56$ predicted segmentation mask, and its first channel is then resized to 7×7 and multiplied onto the original feature map in an element-wise manner. We compute the binary cross entropy loss using the ground-truth face region mask with an extra bitwise inverted channel. The loss weight γ is set to be 0.5.

5. Experiments

We conduct experimental evaluations to show the feasibility of our approach for synthesizing training datasets. We compare our method with existing real datasets and data synthesis approach in terms of gaze estimation accuracy.

5.1. Experimental Settings

MPIIFaceGaze [62] consists of more than 38,000 images of 15 subjects collected from the daily laptop usage. Since we use this dataset only as the source for synthesis, we restricted the source images to be nearly frontal and removed reconstruction failure cases. To ensure the balance between persons as the training data, we randomly downsample or upsample to 1,500 images for each person. **ETH-XGaze** [58] contains 80 subjects of the training set (with published ground truth labels) and 15 subjects of the testing set, with more than 1,000,000 images in total. We use the public evaluation server for evaluating the accuracy on the testing set. **EYEDIAP** [14] consists of more than 4-hours video data captured by VGA and HD cameras, using continuous screen targets or 3D floating object targets. We treated the screen target (CS) and floating target (FT) subsets separately, and sampled one image every 5 frames from the VGA videos following pre-processing by Park *et al.* [33]. **GazeCapture** [27] consists of more than 2 million images crowdsourced from more than 1,300 subjects. We used the metadata provided by Park *et al.* [33] for data normalization. **Gaze360** [26] consists of indoor and outdoor images from 238 subjects with very large head pose and gaze range. We do not consider the cases where eyes are invisible and apply pre-processing following Cheng *et al.* [7], which results in 84,902 images.

To preprocess these datasets, we apply the data normalization scheme commonly used by appearance-based gaze estimation models [59]. Unless otherwise noted, we follow the normalization setting used in ETH-XGaze dataset [58]. The virtual camera’s focal length is set to 960 mm, and the distance from the camera origin to the face center is set to 300 mm. Face images are rendered in 448×448 pixels, and down-scaled to 224×224 pixels before being fed into CNNs. Please note that the normalization process requires head pose information even for test images. We use the landmark locations provided by each dataset and obtain the head pose by fitting a 6-landmark 3D face model using the PnP algorithm [13].

We also directly render our synthetic images in the same normalized camera space. We render face images using the same camera intrinsic and distance parameters, and further apply normalization operation to cancel in-plane rotations if needed.

As the most simple baseline against our mask-guided network, we use a gaze estimation network with the ResNet-50 [23] backbone. This corresponds to the pro-

posed network without the segmentation branch and attention mechanism, and the overall architecture is equivalent to the baseline model used for the evaluation in ETH-XGaze [58].

5.2. Dataset Extrapolation

We first focus on the dataset extrapolation cases where the source MPIIFaceGaze dataset is extended to have a similar head pose distribution as the target ETH-XGaze² and EYEDIAP datasets. We use the head pose values obtained through the data normalization process, and each source image is reconstructed and rendered with 16 new head poses randomly chosen from the target dataset. To avoid extremely large head poses where the eyes are fully occluded, we set a threshold of 80 degrees on the norm of the pitch-yaw vector and discarded such target poses. As a result, the MPIIFaceGaze is extended to three synthetic datasets for ETH-XGaze, EYEDIAP CS, and EYEDIAP FT, respectively, all with 360,000 images. We refer to these datasets as MPII-NV.

We evaluate how our data synthesis approach improves the performance compared to other baseline training data. As a real image baseline, we used the Gaze360 dataset which mostly covers the target gaze range. In addition, we use ST-ED [64] to generate a neural rendering-based synthetic baseline. We used their pre-trained model and we only redirected the head pose of the source dataset in the same manner as our dataset. The dataset is named MPII-NV-STED and its samples are shown in the first two rows of Fig. 7. The head pose and gaze distributions of the source, target, and synthetic datasets are shown in Fig. 6, together with the gaze distribution of Gaze360. Since we synthesized the data based on head pose distribution, it can be seen that the gaze distribution does not exactly match the target, but only roughly overlaps.

Since the pretrained model of ST-ED can only output 128×128 images, we also downscaled the test images when evaluating the model trained on ST-ED. As reported in prior work [58], there is no need to modify the architecture for different image sizes due to the average pooling layer at the end of the ResNet.

The results are summarized in Table 1. The second to fourth rows are the real datasets, and the last three rows are the extended synthetic datasets created by ST-ED [64] and our proposed synthesis method, respectively. All corresponds to the baseline model performance except the last row being trained with our proposed mask-guided model.

The original MPII has the narrowest gaze range and resulted in the highest errors for all test datasets, and our proposed synthetic dataset and model (last row) reduced these errors by 60%, 57%, 61%, and 35% for each test dataset, respectively. ETH-XGaze and Gaze360 both contain wider

²We used the training subset as the target head pose distribution.

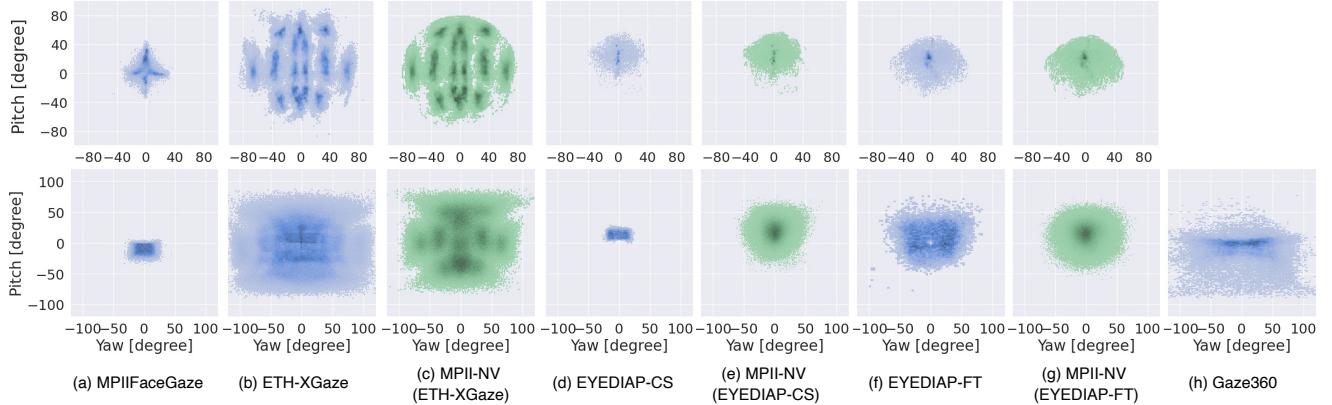


Figure 6. Distributions of head pose (top row) and gaze direction (bottom row). (a) source MPIIFaceGaze, (b) target ETH-XGaze, (c) synthesized dataset by extending MPIIFaceGaze for ETH-XGaze distribution, (d) target EYEDIAP (CS), (e) synthesized dataset for EYEDIAP (CS), (f) target EYEDIAP (FT), (g) synthesized dataset for EYEDIAP (FT), and (h) Gaze360 (head pose not provided).

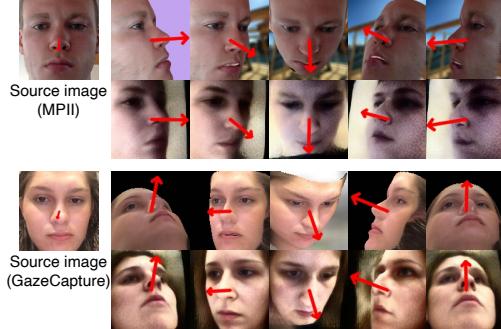


Figure 7. Examples of synthesized images. For each source image, the first row is from our proposed method, and the second row is from ST-ED [64].

| Training \ Test | ETH-XGaze | | EYEDIAP | |
|----------------------|-------------|-------------|------------|-------------|
| | Train | Test | CS | FT |
| MPIIFaceGaze [63] | 32.5 | 33.0 | 14.3 | 24.5 |
| ETH-XGaze Train [58] | - | - | 8.8 | 13.0 |
| Gaze360 [26] | 17.5 | 18.2 | 7.6 | 12.6 |
| MPII-NV-STED [64] | 29.1 | 30.6 | 10.3 | 20.7 |
| MPII-NV | 14.4 | 16.1 | 5.8 | 17.3 |
| MPII-NV (Mask) | 13.0 | 14.3 | 5.6 | 16.0 |

Table 1. Comparison of gaze estimation errors in degree. Each row corresponds to a training dataset, and the second to fifth columns show mean angular errors for each test dataset.

gaze range and perform better on other datasets, but still inferior to our dataset. While MPII-NV-STED has a wide gaze range as our dataset, it does not effectively improve the performance. This indicates the difficulty of maintaining ground-truth gaze labels through neural rendering, while our method faithfully reproduces the authentic gaze direc-

tion by sampling the original appearance. As a result, our method achieved the best performance on ETH-XGaze and EYEDIAP CS. The only exception was the EYEDIAP FT subset, where better performance was obtained when using real data. EYEDIAP FT has a larger offset between gaze and head pose because of the use of physical gaze targets, and our data synthesized based on head pose cannot fully reproduce the target gaze distribution (Fig. 6). However, our method still outperforms the ST-ED result.

Comparison with ST-ED [64] As shown in Fig. 7, ST-ED cannot preserve the same identity as the source image in MPIIFaceGaze because their model was trained on the GazeCapture dataset. To further analyze the difference between our approach and ST-ED, we compare the performances using GazeCapture as the source dataset.

We randomly chose 1,000 out of the 1,374 subjects in GazeCapture, and further randomly chose 30 images from each subject. We sample 12 new head poses from ETH-XGaze for each source image. The gaze estimation error of the baseline model on ETH-XGaze Training set was **20.6** and **26.3** degrees for our synthetic data and ST-ED, respectively. This again proved that the neural rendering approach still cannot provide accurate images and gaze label pairs to be used as training data.

Effect of Head Pose Prior As discussed earlier, the head pose distribution of the target dataset can be obtained from unlabelled images. However, in practice, there may be cases where adaptation must be performed without any image samples from the target environment. To represent such use case scenarios, we further evaluate the performance by synthesizing samples without relying on any prior knowledge about the target dataset. Specifically, we assumed zero-mean normal distribution for both yaw and pitch of the gaze angle, and varied the standard deviation σ from 5 to 40 de-

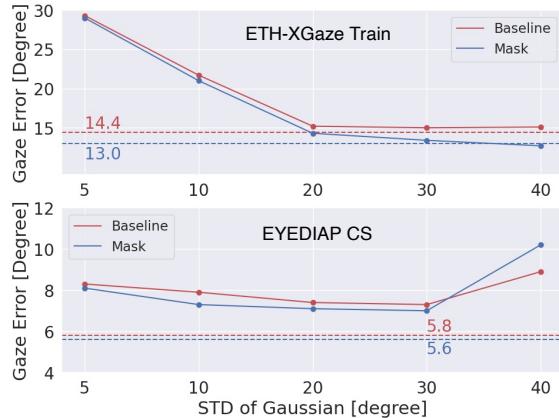


Figure 8. The gaze estimation errors of both baseline and mask-guided models with respect to the variance σ of Gaussian sampling. The horizontal dashed lines correspond to the error reported in Table 1 directly using the target head pose distribution.

rees. Source data was MPIIFaceGaze, and we trained and tested the mask-guided model on ETH-XGaze training set and EYEDIAP.

Figure 8 shows the gaze estimation errors with respect to σ for ETH-XGaze training set and EYEDIAP CS. When σ is very small, the synthetic dataset still cannot cover the gaze distributions of both datasets and the models do not perform well. As σ increases, the coverage of the synthetic dataset also increases and the performance approaches the best case scenario. However, since EYEDIAP has a narrow gaze range (Fig. 6) compared to ETH-XGaze, the errors start to increase after $\sigma = 30$. This indicates that, although synthesizing data over an excessively wide range may adversely affect the performance, sufficient performance can be obtained without prior knowledge on head pose distribution.

5.3. Ablation Studies

We evaluate the effect of data augmentation and mask-guided model using MPII-NV (tested on the ETH-XGaze Train set). We also include the frontal camera of ETH-XGaze Training set as another source dataset to see the upper-bound performance of our approach on the ETH-XGaze Test set. From the frontal image, we synthesized images under head poses corresponding to all of the 18 cameras (XGazeF-NV). In this within-dataset setting, the best case performance using the real ETH-XGaze Train set was **4.5** degrees.

In the first four rows in Table 2, we can observe the performance gain by adding random color, random scene image, and weak-lighting. Single black background tends to overfit, and it is effectively alleviated by adding random colors and random scenes. We keep 40% color background

| Ablations \ Datasets | MPII-NV | XGazeF-NV |
|----------------------|-------------|------------|
| Black | 24.4 | 21.6 |
| + Color (1:1) | 19.6 | 18.7 |
| + Scene (1:1:3) | 14.8 | 12.9 |
| + Weak-light | 14.4 | 11.2 |
| SimGAN [40] | 15.0 | 10.0 |
| DANN [16] | 20.6 | 19.1 |
| PADACO [28] | 16.5 | 28.7 |
| Mask-guided (ours) | 13.0 | 8.3 |

Table 2. Ablation study for analyzing the data augmentation and models. The data augmentation components are evaluated on baseline model.

images to avoid generalizing poorly on some simple background test data. Finally, the increased diversity of lighting made the model more robust.

In addition, we compare the proposed mask-guided model with some existing domain adaptation methods in the last four rows in Table 2. We used our implementations of SimGAN [40], DANN [16], and PADACO [28], all using the baseline gaze estimation network architecture. Overall, these domain adaptation methods failed to perform better than the baseline model (fourth row). In contrast, our mask-guided model effectively reduced the error by benefiting from the synthesis process. The best performance (8.3 degrees) using XGazeF-NV is comparative with the result using the real ETH-XGaze Train set, while indicating the effect of remaining domain gaps.

6. Conclusion

In this work, we presented a novel learning-by-synthesis pipeline for appearance-based full-face gaze estimation. Our approach utilizes 3D face reconstruction to synthesize training datasets with novel head poses, while keeping accurate gaze labels via projective matching. We also proposed the mask-guided gaze estimation model with synthetic data augmentation and showed that our approach effectively extends the gaze target range of the source dataset. Through experiments, our approach achieved better performance than the state-of-the-art neural rendering approach.

All datasets used in this work were collected with IRB approval or consent from participants [14, 26, 27, 58, 62]. Although the proposed method creates synthetic faces, ethical issues are minimum because the identity is generated from the source datasets. In other words, our method cannot extend the diversity of human faces by synthesizing new identities. As discussed in the experiment, it is also difficult for our method to extend the limited head-gaze offset distribution in the source dataset. It is an important future work to explore learning-by-synthesis approaches to cover different data diversity requirements.

References

- [1] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *Proc. FG*, 2018. 4
- [2] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE TPAMI*, 35(12):2930–2940, 2013. 3
- [3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH*, 1999. 3
- [4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proc. ICCV*, 2017. 3
- [5] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proc. AAAI*, 2020. 2
- [6] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proc. ECCV*, 2018. 2
- [7] 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
- [8] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE TIP*, 29:5259–5272, 2020. 2
- [9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proc. CVPRW*, 2019. 3
- [10] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *TOG*, 40(4), jul 2021. 3
- [11] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proc. ECCV*, 2018. 3
- [12] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rtgene: Real-time eye gaze estimation in natural environments. In *Proc. ECCV*, 2018. 1, 2
- [13] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 4, 6
- [14] 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 *Proc. ETRA*, 2014. 2, 6, 8
- [15] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE TPAMI*, 32(8):1362–1376, 2010. 1
- [16] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. ICML*, 2015. 8
- [17] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE TBE*, 53(6):1124–1133, 2006. 2
- [18] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 3ddfa. <https://github.com/cleardusk/3DDFA>, 2018. 3, 4
- [19] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proc. ECCV*, 2020. 3
- [20] Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. Domain adaptation gaze estimation by embedding with prediction consistency. In *Proc. ACCV*, 2020. 2
- [21] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proc. CVPR*, 2016. 1
- [22] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE TPAMI*, 32(3):478–500, 2010. 2
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 5, 6
- [24] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28(5):445–461, 2017. 2
- [25] Amin Jourabloo and Xiaoming Liu. Pose-invariant 3d face alignment. In *Proc. ICCV*, 2015. 3
- [26] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proc. ICCV*, 2019. 1, 2, 6, 7, 8
- [27] Kyle Kraftka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proc. CVPR*, 2016. 1, 2, 3, 6, 8
- [28] Felix Kuhnke and Joern Ostermann. Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces. In *Proc. ICCV*, pages 10163–10172, 2019. 2, 8
- [29] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. In *Proc. SIGGRAPH Asia*, 2017. 3
- [30] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proc. ICCV*, 2021. 2
- [31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. CVPR*, 2015. 5
- [32] Iacopo Masi, Tal Hassner, Anh Tuan Tran, and Gérard Medioni. Rapid synthesis of massive face sets for improved face recognition. In *Proc. FG*, 2017. 3
- [33] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proc. ICCV*, 2019. 2, 6
- [34] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Proc. ECCV*, 2018. 1
- [35] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proc. AVSS*, 2009. 3

- [36] Stylianos Ploumpis, Evangelos Ververas, Eimear O’Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William Smith, Baris Gecer, and Stefanos P Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE TPAMI*, 2020. 3
- [37] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 4
- [38] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proc. ECCV*, 2016. 1
- [39] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proc. CVPR*, 2018. 1
- [40] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proc. CVPR*, 2017. 1, 2, 8
- [41] Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. Gaze locking: Passive eye contact detection for human-object interaction. In *Proc. UIST*, 2013. 2
- [42] Hao Su, Charles R. Qi, Yangyan Li, and Leonidas J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. 2015. 1
- [43] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proc. CVPR*, 2014. 1, 2
- [44] Ryuhei Takahashi, Atsushi Hashimoto, Motoharu Sonogashira, and Masaaki Iiyama. Partially-shared variational auto-encoders for unsupervised domain adaptation with target shift. In *Proc. ECCV*, pages 1–17. Springer, 2020. 2
- [45] Kar-Han Tan, David J Kriegman, and Narendra Ahuja. Appearance-based eye gaze estimation. In *Proc. WACV*, 2002. 2
- [46] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proc. CVPRW*, 2018. 1
- [47] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proc. CVPR*, 2017. 3
- [48] Ulrich Weidenbacher, Georg Layher, P-M Strauss, and Heiko Neumann. A comprehensive head pose and gaze database. In *3rd IET International Conference on Intelligent Environments*, pages 455–458, 2007. 2
- [49] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proc. ETRA*, 2016. 2
- [50] Erroll Wood, Tadas Baltrušaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *Proc. ICCV*, 2015. 1
- [51] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proc. ICCV*, pages 3681–3691, October 2021. 1
- [52] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. A 3D Morphable Model of the Eye Region. In *EG - Posters*, 2016. 2
- [53] Yunyang Xiong, Hyunwoo J Kim, and Vikas Singh. Mixed effects neural networks (menets) with applications to gaze estimation. In *Proc. CVPR*, 2019. 2
- [54] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, 2015. 5
- [55] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proc. CVPR*, 2019. 2
- [56] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Proc. CVPR*, 2020. 2
- [57] Yuxiao Hu, Dalong Jiang, Shuicheng Yan, Lei Zhang, and Hongjiang zhang. Automatic 3d reconstruction for face recognition. In *Proc. FG*, 2004. 3
- [58] 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 *Proc. ECCV*, 2020. 1, 2, 3, 4, 5, 6, 7, 8
- [59] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proc. ETRA*, 2018. 4, 6
- [60] Xucong Zhang, Yusuke Sugano, Andreas Bulling, and Otmar Hilliges. Learning-based region selection for end-to-end gaze estimation. In *Proc. BMVC*, 2020. 2
- [61] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proc. CVPR*, 2015. 1, 2, 3
- [62] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It’s written all over your face: Full-face appearance-based gaze estimation. In *Proc. CVPRW*, 2017. 1, 2, 4, 5, 6, 8
- [63] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE TPAMI*, 41(1):162–175, 2019. 2, 7
- [64] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. In *Proc. NeurIPS*, 2020. 1, 2, 6, 7
- [65] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. 4
- [66] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proc. CVPR*, 2020. 3

- [67] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *Proc. CVPR*, 2016. [3](#)
- [68] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE TPAMI*, 41(1):78–92, 2019. [3](#)
- [69] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *Computer Graphics Forum*, 37(2):523–550, 2018. [3](#)