Few-Shot Adaptive Gaze Estimation

Seonwook Park[12*], Shalini De Mello[1*], Pavlo Molchanov[1], Umar Iqbal[1], Otmar Hilliges[2], Jan Kautz[1]
[1]NVIDIA, [2]ETH Zürich
{spark, otmarh}@inf.ethz.ch; {shalinig, pmolchanov, uiqbal, jkautz}@nvidia.com

Abstract

*Inter-personal anatomical differences limit the accuracy of person-independent gaze estimation networks. Yet there is a need to lower gaze errors further to enable applications requiring higher quality. Further gains can be achieved by personalizing gaze networks, ideally with few calibration samples. However, over-parameterized neural networks are not amenable to learning from few examples as they can quickly over-fit. We embrace these challenges and propose a novel framework for Few-shot Adaptive GaZE Estimation (*FAZE*) for learning person-specific gaze networks with very few (*≤ 9*) calibration samples.* FAZE *learns a rotation-aware latent representation of gaze via a disentangling encoder-decoder architecture along with a highly adaptable gaze estimator trained using meta-learning. It is capable of adapting to any new person to yield significant performance gains with as few as* 3 *samples, yielding state-of-the-art performance of* 3.18° *on GazeCapture, a 19% improvement over prior art. We open-source our code at* https://github.com/NVlabs/few_shot_gaze [1].

1. Introduction

Estimation of human gaze has numerous applications in human-computer interaction [7], virtual reality [28], automotive [41] and content creation [46], among others. Many of these applications require high levels of accuracy (cf. [3, 37, 15, 2]). While appearance-based gaze estimation techniques that use Convolutional Neural Networks (CNN) have significantly surpassed classical methods [51] for in-the-wild settings, there still remains a significant gap towards applicability in high-accuracy domains. The currently lowest reported person-independent error of 4.3° [6] is equivalent to 4.7cm at a distance of 60cm, which restricts use of such techniques to public display interactions [54] or crowd-sourced attention analysis [24].

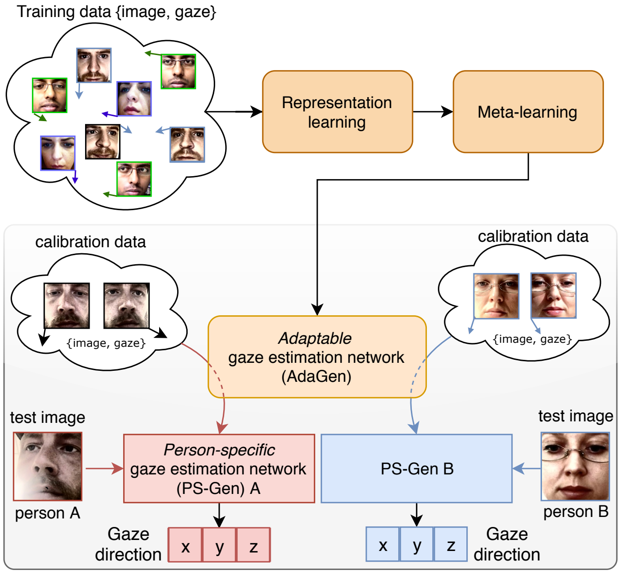


Figure 1: Overview of the FAZE framework. Given a set of training images with ground-truth gaze direction information, we first learn a latent feature representation, which is tailored specifically for the task of gaze estimation. Given the features, we then learn an adaptable gaze estimation network, AdaGEN, using meta-learning which can be adapted easily to a robust person-specific gaze estimation network (PS-GEN) with very little calibration data.

High-accuracy gaze estimation from images is difficult because it requires either explicit or implicit fitting of a person-specific eye-ball model to the image data and the estimation of their visual and optical axes. Moreover, it is well understood that inter-subject anatomical differences affect gaze estimation accuracy [9]. Classical model-based techniques can often be personalized via few (9 or less) samples (e.g., [9, 11]) but are not robust to image variations in uncontrolled settings. While feasible, subject-specific training of CNNs requires thousands of samples and is clearly impractical [53]. Few-shot personalization of CNNs is difficult because training of highly overparametrized models with only few training samples will lead to over-fitting.

[*] The first two authors contributed equally.

[1] This includes a real-time demo which takes < 10 seconds to record 9 calibration points for a new user and ~ 1 minute to train a personalized network on a laptop with an NVIDIA GTX GeForce 1060 GPU.

We tackle these many-fold challenges by proposing FAZE, a framework for learning gaze estimation networks for new subjects using very few calibration samples (Fig. 1). It consists of: i) learning a rotation-aware latent representation of gaze via a disentangling transforming encoder-decoder architecture ii) with these features learning a highly adaptable gaze estimator using meta-learning, and iii) adapting it to any new person to yield significant performance gains with as few as 3 samples.

In order to learn a robust representation for gaze, we take inspiration from recent work on transforming encoder-decoder architectures [12, 47] and design a rotation-equivariant pair of encoder-decoder functions. We disentangle the factors of appearance, gaze and head pose in the latent space and enforce equivariance by decoding explicitly rotated latent codes to images of the *same* person but with a *different* gaze direction compared to the input (via a ℓ_1 reconstruction loss). The equivariance property of our gaze representation further allows us to devise a novel *embedding consistency* loss term that further minimizes the intra-person differences in the gaze representation. We then leverage the proposed latent embedding to learn person-specific gaze estimators from few samples. To this end we use a meta-learning algorithm to learn *how to learn* such estimators. We take inspiration from the recent success of meta-learning [1] for few-shot learning in several other vision tasks [5, 10, 25]. To the best of our knowledge, we are the first to cast few-shot learning of person-specific gaze estimators as one of multi-task learning where each subject is seen as a new task in the meta-learning sense.

We evaluate the proposed framework on two benchmark datasets and show that our meta-learned network with its latent gaze features can be successfully adapted with very few ($k \leq 9$) samples to produce accurate person-specific models. We demonstrate the validity of our design choices with detailed empirical evidence, and demonstrate that our proposed framework outperforms state-of-the-art methods by significant margins. In particular, we demonstrate improvements of 13% ($3.94° \rightarrow 3.42°$) on the MPIIGaze dataset, and 19% ($3.91° \rightarrow 3.18°$) on the GazeCapture dataset over the approach of [20] using just 3 calibration samples.

To summarize, the main contributions of our work are:

- FAZE, a novel framework for learning person-specific gaze networks with few calibration samples, fusing the benefits of rotation-equivariant feature learning and meta-learning.
- A novel encoder-decoder architecture that disentangles gaze direction, head pose and appearance factors.
- A novel and effective application of meta-learning to the task of few-shot personalization.
- State-of-the-art performance ($3.14°$ with $k = 9$ on MPIIGaze), with consistent improvements over existing methods for $1 \leq k \leq 256$.

2. Related Work

Gaze Estimation. Appearance-based gaze estimation [40] methods that map images directly to gaze have recently surpassed classical model-based approaches [11] for in-the-wild settings. Earlier approaches in this direction assume images captured in restricted laboratory settings and use direct regression methods [22, 21] or learning-by-synthesis approaches combined with random forests to separate head-pose clusters [39]. More recently, the availability of large scale datasets such as MPIIGaze [51] and GazeCapture [17], and progress in CNNs have rapidly moved the field forward. MPIIGaze has become a benchmark dataset for in-the-wild gaze estimation. For the competitive person-independent within-MPIIGaze leave-one-person-out evaluation, gaze errors have progressively decreased from $6.3°$ for naively applying a LeNet-5 architecture to eye-input [51] to the current best of $4.3°$ with an ensemble of multi-modal networks based on VGG-16 [6]. Proposed advancements include the use of more complex CNNs [53]; more meaningful use of face [52, 17] and multi-modal input [17, 6, 48]; explicit handling of differences in the two eyes [4]; greater robustness to head pose [55, 30]; improvements in data normalization [49]; learning more informed intermediate representations [26]; using ensembles of networks [6]; and using synthetic data [36, 45, 19, 27, 30].

However, person-independent gaze errors are still insufficient for many applications [3, 37, 15, 2]. While significant gains can be obtained by training person-specific models, it requires many thousands of training images per subject [53]. On the other hand, CNNs are prone to over-fitting if trained with very few ($k \leq 9$) samples. In order to address this issue, existing approaches try to adapt person-independent CNN-based features [17, 27] or points-of-regard (PoR) [50] to person-specific ones via hand-designed heuristic functions. Some methods also train a Siamese network with pairs of images of the same subject [20].

Learned Equivariance. Generalizing models learned for regression tasks to new data is a challenging problem. However, recent works show improvements from enforcing the learning of equivariant mappings between input, latent features, and label spaces [13, 33], via so-called transforming encoder-decoder architectures [12]. In [47], this idea is expanded to learn complex phenomena such as the orientation of synthetic light sources and in [33] the method is applied to real-world multi-view imagery to improve semi-supervised human pose estimation. In contrast, we learn from very noisy real-world data while successfully disentangling the two noisily-labeled phenomena of gaze direction and head orientation.

Few-shot Learning. Few-shot learning aims to learn a new task with very few examples [18]. This is a non-trivial problem for highly over-parameterized deep networks as it leads to over-fitting. Recently, several promising meta-learning [38, 44, 32, 34, 5, 23, 31] techniques have been proposed that learn *how to learn* unique but similar tasks in a few-shot manner using CNNs. They have been shown to be successful for various few-shot visual learning tasks including object recognition [5], segmentation [29], viewpoint estimation [42] and online adaptation of trackers [25]. Inspired by their success, we use meta-learning to learn how to learn person-specific gaze networks from few examples. To the best of our knowledge we are the first to cast person-specific gaze estimation as a multi-task problem in the context of meta-learning, where each subject is seen as a new task for the meta-learner. Our insight is that meta-learning lends itself well to few-shot gaze personalization and leads to performance improvements.

3. Method

In this section, we describe how we perform gaze estimation from challenging in-the-wild imagery, with minimal burden to the user. The latter objective can be fulfilled by designing our framework to adapt well even with very few calibration samples ($k \leq 9$). We first provide an overview of the FAZE framework and its three stages.

3.1. The FAZE framework

We design FAZE (Fig. 1) with the understanding that a person-specific gaze estimator must encode factors particular to the person, yet at the same time, leverage insights from observing the eye-region appearance variations across a large number of people with different head pose and gaze direction configurations. The latter is important for building models that are robust to extraneous factors such as poor image quality. Thus, the first step in FAZE is to learn a generalizable latent embedding space that encodes information pertaining to the gaze-direction, including person-specific aspects. We detail this in Sec. 3.2.

Provided that good and consistent features can be learned, we can leverage meta-learning to learn how to learn few-shot person-specific gaze estimators for these features. This results in few-shot learners which generalize better to new people (tasks) without overfitting. Specifically, we use the MAML meta-learning algorithm [5]. For our task, MAML learns a set of initial network weights which allow for fine-tuning without the usual over-fitting issues that occur with low k. Effectively, it produces a highly Adaptable Gaze Estimation Network (AdaGEN). The final step concerns the adaptation of MAML-initialized weights to produce person-specific models (PS-GEN) for each user. We describe this in Sec. 3.3.

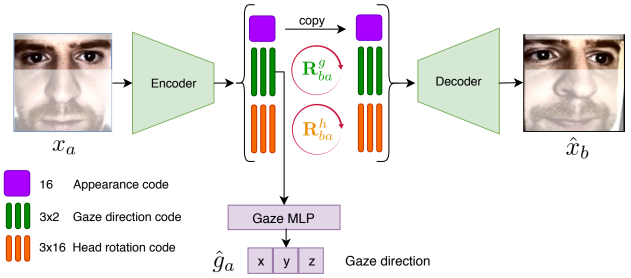


Figure 2: Disentangling appearance, gaze and head pose variations from an image with our Disentangling Transforming Encoder-Decoder (DT-ED). We learn to translate between pairs of images of the same person by rotating the gaze and head pose codes. The encoder-decoder are supervised by a pixel-wise L_1 loss (Eq. 3), with the gaze embedding additionally supervised via gaze regression (Eq. 5).

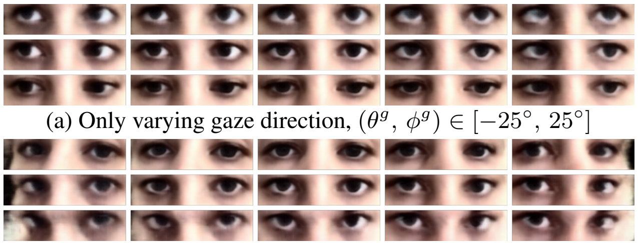
3.2. Gaze-Equivariant Feature Learning

In this section, we explain how the learning of a function, which understands equivalent rotations in input data and output label can lead to better generalization in the context of our final task of person-specific gaze estimation. In addition, we: (a) show how to disentangle eyeball and head rotation factors leading to better distillation of gaze information, and (b) introduce a frontalized *embedding consistency* loss term to specifically aid in learning consistent frontal gaze codes for a particular subject.

3.2.1 Architecture Overview

In learning a generalizable latent embedding space representing gaze, we apply the understanding that a relative change in gaze direction is easier to learn in a person-independent manner [20]. We extend the transforming encoder-decoder architecture [12, 47] to consider three distinct factors apparent in our problem setting: gaze direction, head orientation, and other factors related to the appearance of the eye region in given images (Fig. 2). We disentangle the three factors by *explicitly* applying separate and known differences in rotations (eye gaze and head orientation) to the respective sub-codes. We refer to this architecture as the Disentangling Transforming Encoder-Decoder (DT-ED).

For a given input image $\mathbf{x}$, we define an encoder $\mathcal{E} : \mathbf{x} \rightarrow \mathbf{z}$ and a decoder $\mathcal{D} : \mathbf{z} \rightarrow \hat{\mathbf{x}}$ such that $\mathcal{D}\left(\mathcal{E}(\mathbf{x})\right) = \hat{\mathbf{x}}$. We consider the latent space embedding $\mathbf{z}$ as being formed of 3 parts representing: appearance ($\mathbf{z}^a$), gaze direction or eyeball rotation ($\mathbf{z}^g$), and head pose ($\mathbf{z}^h$), which can be expressed as: $\mathbf{z} = \left\{\mathbf{z}^a; \mathbf{z}^g; \mathbf{z}^h\right\}$ where gaze and head codes are flattened to yield a single column. We define $\mathbf{z}^g$ as having dimensions $(3 \times F^g)$ and $\mathbf{z}^h$ as having dimensions $\left(3 \times F^h\right)$ with $F \in \mathbb{N}$. With these dimensions, it is possible to apply a rotation matrix to explicitly rotate these $3D$ latent space embeddings using rotation matrices.



(a) Only varying gaze direction, $(\theta^g,\, \phi^g) \in [-25°,\, 25°]$

(b) Only varying head orientation, $\left(\theta^h,\, \phi^h\right) \in [-20°,\, 20°]$

Figure 3: Our disentangled rotation-aware embedding spaces for gaze direction and head pose are demonstrated by predicting embeddings $\hat{\mathbf{z}}^g$, $\hat{\mathbf{z}}^h$ from a given sample, rotating it to 15 points each, and then decoding them.

The frontal orientation of eyes and heads in our setting can be represented as $(0,\, 0)$ in Euler angles notation for azimuth and elevation, respectively assuming no roll, and using the $x - y$ convention. Then, the rotation of the eyes and the head from the frontal orientation can be described using (θ^g, ϕ^g) and $\left(\theta^h, \phi^h\right)$ in Euler angles and converted to rotation matrices defined as,

$$\mathbf{R}^{(\theta,\, \phi)} = \begin{bmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}. \tag{1}$$

While training DT-ED, we input a pair of images of a person $\mathbf{x}_a$ and $\mathbf{x}_b$. We can calculate $\mathbf{R}^g_{ba} = \mathbf{R}^g_b \left(\mathbf{R}^g_a\right)^{-1}$ to describe the change in gaze direction in going from sample a to sample b of the same person. Likewise for head rotation, $\mathbf{R}^h_{ba} = \mathbf{R}^h_b \left(\mathbf{R}^h_a\right)^{-1}$. This can be done using the *ground-truth* labels for gaze ($\mathbf{g}_a$ and $\mathbf{g}_b$) and head pose ($\mathbf{h}_a$ and $\mathbf{h}_b$) for the pair of input samples. The rotation of the latent code $\mathbf{z}^g_a$ can then be expressed via the operation $\hat{\mathbf{z}}^g_b = \mathbf{R}^g_{ab}\mathbf{z}^g_a$. At training time, we enforce this code to be equivalent to the one extracted from image $\mathbf{x}_b$, via a reconstruction loss (Eq. 3). We assume the rotated codes $\hat{\mathbf{z}}^h_b$ and $\hat{\mathbf{z}}^g_b$, along with the appearance-code $\mathbf{z}^a_a$, to be sufficient for reconstructing $\mathbf{x}_b$ through the decoder function such that, $\mathcal{D}\left(\hat{\mathbf{z}}_b\right) = \mathbf{x}_b$. More specifically, given the encoder output $\mathcal{E}\left(\mathbf{x}_a\right) = \mathbf{z}_a = \left\{\mathbf{z}^a_a;\, \mathbf{z}^g_a;\, \mathbf{z}^h_a\right\}$, we assume the rotated version of $\mathbf{x}_a$ to match the embedding of $\mathbf{x}_b$, that is we assume $\left\{\hat{\mathbf{z}}^a_b;\, \hat{\mathbf{z}}^g_b;\, \hat{\mathbf{z}}^h_b\right\} = \left\{\mathbf{z}^a_a;\, \mathbf{R}^g_{ba}\mathbf{z}^g_a;\, \mathbf{R}^h_{ba}\mathbf{z}^h_a\right\}$ (See Fig. 2).

This approach indeed applies successfully to noisy real-world imagery, as shown in Fig. 3 where we map a sample into the gaze and head pose latent spaces, rotate to the frontal orientation, and then again rotate by a pre-defined set of 15 yaw and pitch values and reconstruct the image via the decoder. We can see that the factors of gaze direction and head pose are fully disentangled and DT-ED succeeds in the challenging task of eye-region frontalization and redirection from monocular RGB input.

We train the FAZE transforming encoder-decoder architecture using a multi-objective loss function defined as,

$$\mathcal{L}_\text{full} = \lambda_\text{recon}\mathcal{L}_\text{recon} + \lambda_\text{EC}\mathcal{L}_\text{EC} + \lambda_\text{gaze}\mathcal{L}_\text{gaze}, \tag{2}$$

where we empirically set $\lambda_\text{recon} = 1$, $\lambda_\text{EC} = 2$, and $\lambda_\text{gaze} = 0.1$. The individual loss terms are explained in the following sub-sections.

3.2.2 Reconstruction Loss

To guide learning of the encoding-decoding process, we apply a simple ℓ_1 reconstruction loss. Given an input image $\mathbf{x}_b$ and reconstructed $\hat{\mathbf{x}}_b$ obtained by decoding the rotated embeddings $\hat{\mathbf{z}}_b$ of image $\mathbf{x}_a$, the loss term is defined as,

$$\mathcal{L}_\text{recon}\left(\mathbf{x}_b,\, \hat{\mathbf{x}}_b\right) = \frac{1}{|\mathbf{x}_b|} \sum_{u \in \mathbf{x}_b, \hat{u} \in \hat{\mathbf{x}}_b} |\hat{u} - u|\,, \tag{3}$$

where u and $\hat{u}$ are pixels of images $\mathbf{x}_b$ and $\hat{\mathbf{x}}_b$ respectively.

3.2.3 Embedding consistency Loss

We introduce a novel embedding consistency term, which ensures that the encoder network always embeds images of a person with different appearance but identical gaze direction to similar features. Usually this would require paired images with only gaze directions changed. However, it is intractable to collect such data in the real world, so we instead exploit the learned equivariance of DT-ED. Before measuring the consistency between latent gaze features from different samples, we first frontalize them by applying the inverse of the rotation matrix $\mathbf{R}^g_a$ using ground-truth gaze direction $\mathbf{g}_a$. It should be noted that naively enforcing all gaze features to be similar across persons may disregard the inter-subject anatomical differences which should result in different latent embeddings. Hence, we apply the embedding consistency to intra-subject pairs of images only. We validate this choice through experiments in Sec. 5.1.

Given a batch of B image samples during training, we formally compute the *embedding consistency* loss using,

$$\mathcal{L}_\text{EC} = \frac{1}{B} \sum_{i=1}^{B} \max_{\substack{j=1\ldots B \\ id(i)=id(j)}} d\left(f(\mathbf{z}^g_i), f(\mathbf{z}^g_j)\right), \tag{4}$$

where $f(\mathbf{z}^g) = \left(\mathbf{R}^g\right)^{-1}\mathbf{z}^g$ corresponds to frontalized latent gaze features. The function $id(i)$ provides the person-identity of the i-th sample in the batch, and d is a function based on mean column-wise angular distance (between 3D vectors). The $\max$ function minimizes differences between intra-person features that are furthest apart, and is similar to the idea of batch-hard online triplet mining [35].

During training, we linearly increase λ_EC from 0 until sufficient mini-batches to cover 10^6 images have been processed, to allow for more natural embeddings to arise before applying consistency.

3.2.4 Gaze Direction Loss

Lastly, we add the additional objective of gaze estimation via $\mathcal{G}: \mathbf{z}^g \rightarrow \hat{\mathbf{g}}$, parameterized by a simple multi-layer perceptron (MLP). The gaze direction loss is calculated using,

$$\mathcal{L}_{\text{gaze}}(\hat{\mathbf{g}}, \mathbf{g}) = \arccos\left(\frac{\hat{\mathbf{g}} \cdot \mathbf{g}}{\|\hat{\mathbf{g}}\|\|\mathbf{g}\|}\right). \tag{5}$$

3.3. Person-specific Gaze Estimation

Having learned a robust feature extractor, which is tailored specifically for gaze estimation, our final goal is to learn a personalized gaze estimator with as few calibration samples as possible. A straightforward solution for doing this is to train a person-independent model with the training data used to train DT-ED and simply fine-tune it with the available calibration samples for the given subject. However, in practical setups where only a few calibration samples are available, this approach can quickly lead to overfitting (see experiments in Fig. 4a). In order to alleviate this problem, we propose to use the meta-learning method MAML [5], which learns a highly adaptable gaze network (AdaGEN).

Adaptable Gaze Estimator (AdaGEN) Training. We wish to learn weights θ^* for the AdaGEN gaze prediction model $\mathcal{M}$ such that it becomes a successful few-shot learner. In other words, when $\mathcal{M}_{\theta^*}$ is fine-tuned with only a few "calibration" examples of a new person $\mathcal{P}$ not present in the training set, it can generalize well to unseen "validation" examples of the same person. We achieve this by training it with the MAML meta learning algorithm adapted for few-shot learning.

In conventional CNN training the objective is to minimize the training loss for all the examples of all training subjects. In contrast, for few-shot learning, MAML explicitly minimizes the *generalization* loss of a network *after* minimizing its training loss for a few examples of a particular subject via a standard optimization algorithm, e.g., stochastic gradient descent (SGD). Additionally, MAML repeats this procedure for all subjects in the training set and hence learns from several closely related "tasks" (subjects) to become a successful few shot learner for any new unseen task (subject). We identify that person-specific factors may have few parameters, with only slight but important variations across people such that all people constitute a set of closely related tasks. Our insight is that meta-learning lends itself well to such a problem of personalization.

The overall procedure of meta-learning to learn the optimal θ^* is as follows. We divide the entire set of persons $\mathcal{S}$ into meta-training ($\mathcal{S}^{train}$) and meta-testing ($\mathcal{S}^{test}$) subsets of non-overlapping subjects. During each meta-training iteration n, we randomly select one person $\mathcal{P}^{train}$ from $\mathcal{S}^{train}$ and create a meta-training sample for the person (via random sampling), defined as $\mathcal{P}^{train} = \{\mathcal{D}_c^{train}, \mathcal{D}_v^{train}\}$, containing a calibration set $\mathcal{D}_c^{train} = \{(\mathbf{z^g}_i, \mathbf{g}_i)|i = 1 \ldots k\}$ of k training examples, and a validation set $\mathcal{D}_v^{train} = \{(\mathbf{z^g}_j, \mathbf{g}_j)|j = 1 \ldots l\}$ of another l examples for the same person. Here, $\mathbf{z^g}$ and g refer to the latent gaze representation learned by DT-ED and the ground-truth 3D gaze vector, respectively. Both k and l are typically small (≤ 20) and k represents the "shot" size used in few-shot learning.

The first step in the meta-learning procedure is to compute the loss for the few-shot calibration set $\mathcal{D}_c^{train}$ and update the weights θ_n at step n via one (or more) gradient steps and a learning rate α as,

$$\theta_n' = f(\theta_n) = \theta_n - \alpha \nabla \mathcal{L}^c_{\mathcal{P}^{train}}(\theta_n). \tag{6}$$

With the updated weights θ_n', we then compute the loss for the validation set $\mathcal{D}_v^{train}$ of the subject $\mathcal{P}^{train}$ as $\mathcal{L}^v_{\mathcal{P}^{train}}(\theta_n') = \mathcal{L}^v_{\mathcal{P}^{train}}(f(\theta_n))$ and its gradients w.r.t the initial weights of the network θ_n at that training iteration n. Lastly, we update θ_n with a learning rate of η to explicitly minimize the validation loss as,

$$\theta_{n+1} = \theta_n - \eta \nabla \mathcal{L}^v_{\mathcal{P}^{train}}(f(\theta_n)). \tag{7}$$

We repeat these training iterations until convergence to get the optimal weights θ^*.

Final Person-specific Adaptation. Having learned our encoder and our optimal few-shot person-specific learner $\mathcal{M}_{\theta^*}$, we are now well poised to produce person-specific models for each new person $\mathcal{P}^{test}$ from $\mathcal{S}^{test}$. We fine-tune $\mathcal{M}_{\theta^*}$ with the k calibration images $\mathcal{D}_c^{test}$ to create a personalized model for $\mathcal{P}^{test}$ as

$$\theta_{\mathcal{P}^{test}} = \theta^* - \alpha \nabla \mathcal{L}^c_{\mathcal{P}^{test}}(\theta^*), \tag{8}$$

and test the performance of the personalized model $\mathcal{M}_{\theta_{\mathcal{P}^{test}}}$ on person $\mathcal{P}^{test}$'s validation set $\mathcal{D}_v^{test}$.

4. Implementation Details

4.1. Data pre-processing

Our data pre-processing pipeline is based on [49], a revision of the data normalization procedure introduced in [39]. In a nutshell, the data normalization procedure ensures that a common virtual camera points at the same reference point in space with the head upright. This requires the rotation, tilt, and forward translation of the virtual camera. Please refer to [49] for a formal and complete description, and our supplementary materials for a detailed list of changes.

4.2. Neural Network Configurations

DT-ED. The functions $\mathcal{E}$ and $\mathcal{D}$ in our transforming encoder-decoder architecture can be implemented with any CNN architecture. We select the DenseNet architecture [14] both for our DT-ED as well as for our re-implementation of state-of-the-art person-specific gaze estimation methods [20, 50]. The latent codes $\mathbf{z}_a$, $\mathbf{z}_g$, and $\mathbf{z}_h$ are defined to have dimensions (64), (3×2), and (3×16) respectively. Please refer to supplementary materials for further details.

Gaze MLP. Our gaze estimation function $\mathcal{G}$ is parameterized by a multi-layer perceptron with 64 hidden layer neurons and SELU [16] activation. The MLP outputs 3-dimensional unit gaze direction vectors.

4.3. Training

DT-ED. Following [8], we use a batch size of 1536 and apply linear learning rate scaling and ramp-up for the first 10^6 training samples. We use NVIDIA's Apex library[2] for mixed-precision training. and train our model for 50 epochs with a base learning rate of 5×10^{-5}, l_2 weight regularization of 10^{-4}, and use instance normalization [43].

Gaze MLP. During meta-learning, we use $\alpha = 10^{-5}$ with SGD (Eq. 6), and $\eta = 10^{-3}$ (Eq. 7) with the Adam optimizer (α and β in [5]), and do 5 updates per inner loop iteration. For sampling $\mathcal{D}_v^{train}$ we set $l = 100$. During standard eye-tracker calibration, one cannot assume the knowledge of extra ground-truth beyond the k samples. Thus, we perform the final fine-tuning operation (Eq. 8) for 1000 steps for all values of k and for all people.

4.4. Datasets

GazeCapture [17] is the largest available in-the-wild gaze dataset. We mined camera intrinsic parameters from the web for the devices used, and applied our pre-processing pipeline (Sec. 4.1) to yield input images. For training the DT-ED as well as for meta-learning, we use data from 993 people in the training set specified in [17], each with 1766 samples, on average, for a total of $1.7M$ samples. To ensure within-subject diversity of sampled image-pairs at training time, we only select subjects with $\geq$ 400 samples. For computing our final evaluation metric, we use the last 500 entries from 109 subjects that have at least 1000 samples each. We select the k-shot samples for meta-training and fine-tuning randomly from the remaining samples.

MPIIGaze [51] is the most established benchmark dataset for in-the-wild gaze estimation. In comparison to GazeCapture it has higher within-person variations in appearance including illumination, make-up, and facial hair changes, potentially making it more challenging. We use the images specified in the MPIIFaceGaze subset [52] only for evaluation purposes. The MPIIFaceGaze subset consists of 15 subjects each with 2500 samples on average. We reserve the last 500 images of each subject for final evaluations as is done in [53] and select k calibration samples for personalization by sampling randomly from the remaining samples.

5. Results

For all methods, we report person-specific gaze estimation errors for a range of k calibration samples. For each data point, we perform the evaluation 10 times using k randomly chosen calibration samples. Each evaluation or trial yields a mean gaze estimation error in degrees over all subjects in the test set. The mean error over all such trials is plotted, with its standard deviation represented by the shaded areas above and below the curves. The values at $k = 0$ are determined via $\mathcal{G}(\mathbf{z}^g)$. We train this MLP on the GazeCapture training subset, without any person-specific adaptation.

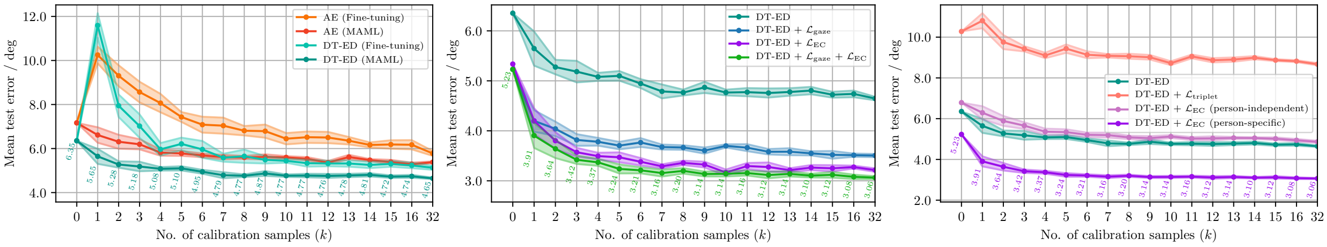
5.1. Ablation Study

We evaluate our method under different settings to better understand the impact of our various design choices. For all experiments, we train the models using the GazeCapture dataset's training set and test on the MPIIGaze dataset. This challenging experiment allows us to demonstrate the generalization capability of our approach across different datasets. The ablation studies are summarized in Fig. 4. We provide additional plots of the results of this ablation study on the test partition of the GazeCapture dataset in the supplementary material.

MAML vs. Finetuning. In Fig. 4a, we first evaluate the impact of meta-learning a few-shot person-adaptive gaze estimator using MAML (Sec. 3.3) and compare its performance with naive finetuning. When no person-specific adaptation is performed (i.e., $k = 0$), the person-independent baseline model $\mathcal{G}(\mathbf{z}^g)$ with the features learned using a vanilla autoencoder (AE) results in a mean test error of $7.17°$. Using MAML for person-specific adaptation with only one calibration sample decreases the error to $6.61°$. The error reduces further as we increase k and reaches a mean value of $5.38°$ for $k = 32$. In contrast, naively finetuning (AE-Finetuning) the gaze estimator results in severe over-fitting and yields very high test errors, in particular, for very low k values. In fact, for $k \leq 3$, the error increases to above the person-independent baseline model. Since the model initialized with weights learned by MAML clearly outperforms vanilla finetuning, in the rest of this section, we always use MAML unless specified otherwise.

Impact of feature representation. Fig. 4a also evaluates the impact of the features used to learn the gaze estimation model. Our proposed latent gaze features (Sec. 3.2) signif-

[2] https://github.com/NVIDIA/apex



(a) (b) (c)

Figure 4: **Ablation Study:** Impact of (a) learning the few-shot gaze estimator using MAML (Sec. 3.3) and using the transforming encoder-decoder for feature learning (Sec. 3.2); (b) different loss terms in Eq. (2) for training the transforming encoder-decoder; and (c) comparison of the different variants of embedding consistency loss term (Eq. (4)). We provide additional results for the test partition of the GazeCapture dataset in the supplementary material.

icantly decrease the error, e.g., $4.87°$ vs. $5.62°$ with $k = 9$ for DT-ED (MAML) and AE (MAML), respectively. Note that the gain remains consistent across all values of k. The only difference between DT-ED and AE is that the latent codes are rotated in DT-ED before decoding. Despite this more difficult task, the learned code clearly better informs the final task of person-specific gaze estimation, showing that disentangling gaze, head pose, and appearance is importance for gaze estimation.

Contribution of loss terms. We evaluate the impact of each loss term described in Eq. (2) (Sec. 3.2) by incorporating them one at a time into the total loss used to train DT-ED. Fig. 4b summarizes the results. Using only the image reconstruction loss $\mathcal{L}_{\text{recon}}$ in Eq. (3), the learned latent gaze features result in an error of $4.87°$ at $k = 9$. Incorporating gaze supervision $\mathcal{L}_{\text{gaze}}$ in Eq. (5) to obtain features that are more informed of the ultimate task of gaze-estimation gives an improvement of 26% from $4.87°$ to $3.60°$. Adding the person-specific embedding consistency term $\mathcal{L}_{\text{EC}}$ in Eq. (4) to $\mathcal{L}_{\text{recon}}$ also reduces the error significantly from $4.87°$ to $3.32°$ at $k = 9$ (an improvement of over 30%). Finally, combining all loss terms improves the performance even further to $3.14°$ (in total, an improvement of 36%).

Analysis of embedding consistency. In order to validate our choice of the embedding consistency loss, in Fig. 4c, we compare its performance with two other possible variants. As described in Sec. 3.2.3, the embedding consistency loss term minimizes the intra-person differences of the frontalized latent gaze features. The main rationale behind this is that the gaze features for a unique person should be consistent while they can be different across subjects due to inter-subject anatomical differences. We further conjecture that preserving these inter-personal differences as opposed to trying to remove them by learning *person-invariant* embeddings is indeed important to obtaining high accuracy for gaze estimation. In order to validate this observation, we introduce a person-*independent* embedding consistency term which also minimizes the inter-person latent gaze feature differences. As is evident from Fig. 4c, enforcing person-independent embedding consistency of the latent gaze features results in increased mean errors. In fact it performs worse than only using the reconstruction loss ($\mathcal{L}_{\text{recon}}$).

One may argue the complete opposite i.e., the latent gaze features should be hugely different for every subject for the best possible subject-specific accuracy, but we did not find this to be the case. To demonstrate this, we apply a triplet loss ($\mathcal{L}_{\text{triplet}}$) [35], which explicitly *maximizes* the inter-personal differences in gaze features in addition to minimizing the intra-person ones. As is evident from Fig 4c this results in a significant increase in the error. This suggests that perhaps factors that quantify the overall appearance of a person's face and define their unique identity may not necessarily be correlated to the anatomical properties that define "person-uniqueness" for the task of gaze estimation.

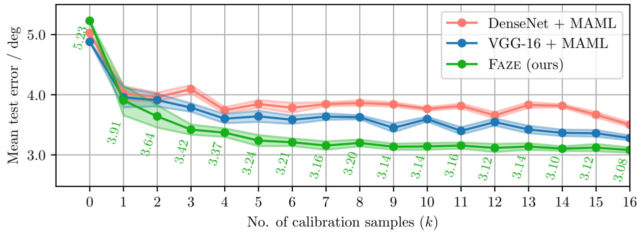
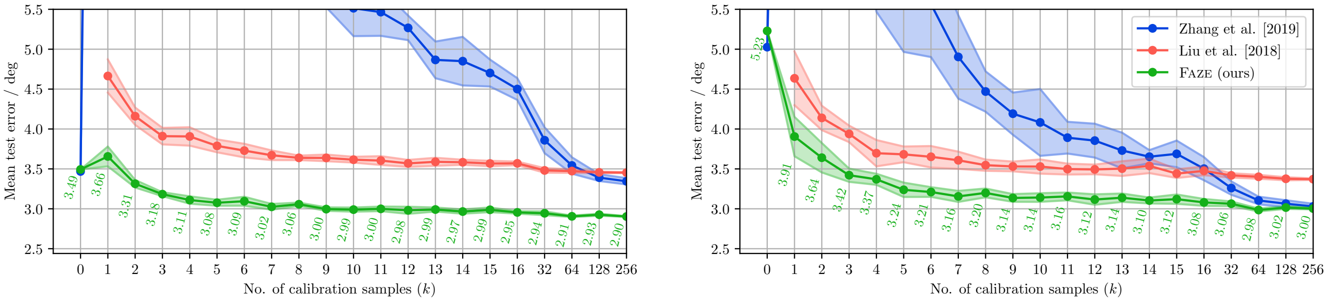


Figure 5: Comparison of FAZE against competitive CNN + MAML baselines, evaluated on MPIIGaze.

5.2. Comparison with CNN + Meta-Learning

An alternative baseline to FAZE can be created by replacing the DT-ED with a standard CNN architecture. We take an identically configured DenseNet (to FAZE) and a VGG-16 architecture for the convolutional layers, then add 2 fully-connected layers each with 256 neurons and train the networks with the gaze objective (Eq. 5). The output of the convolutional layers are used as input to a gaze estimation network trained via MAML to yield the results in Fig. 5. Having been directly trained on the (cross-person) gaze estimation objective, it is expected that the encoder network would make better use of its model capacity as it does not have to satisfy a reconstruction objective. Thus, we can



(a) GazeCapture (test)

(b) MPIIGaze

Figure 6: Comparison of FAZE against state-of-the-art person-specific gaze estimation methods [20, 50]

call these highly competitive baselines. FAZE outperforms these baselines with statistical significance, demonstrating that the DT-ED training and our loss terms yield features which are more amenable to meta-learning, and thus to the final objective of personalized gaze estimation.

5.3. Comparison with State-of-the-Art

Few-shot personalization of CNN models in the context of gaze estimation for very low k is very challenging. Two recent approaches [50, 20] are the most relevant in this direction, and we provide evaluations on highly competitive re-implementations. Our results are presented in Fig. 6 for both the test partition of the GazeCapture dataset and the MPIIGaze dataset. Overall, we show statistically significantly better mean errors over the entire range of $1 \leq k \leq 256$ than all the existing state-of-the-art methods. In addition, our performance between trials is more consistent as shown by the narrower error bands. This indicates robustness to the choice of the k calibration samples.

Ours vs Polynomial fit to PoR. In [50], Zhang et al. fit a 3rd order polynomial function to correct initial point-of-regard (PoR) estimates from a person-independent gaze CNN. To re-implement their method, we train a DenseNet CNN (identical to FAZE) and intersect the predicted gaze ray (defined by gaze origin and direction in 3D with respect to the original camera) with the $z = 0$ plane to estimate the initial PoR and later refine it with a person-specific 3rd order polynomial function. Though this approach performs respectably with $k = 9$, yielding $4.19°$ on MPIIGaze (Fig. 6b), it suffers with lower k especially on GazeCapture. Nonetheless, its performance converges to our performance at $k \geq 128$ showing its effectiveness at higher k despite its apparent simplicity.

Ours vs Differential Gaze Estimation. Liu et al. [20] introduce a CNN architecture for learning to estimate the differences in the gaze yaw and pitch values between pairs of images of the same subject. That is, in order to estimate the gaze their network always requires one *reference* image of a subject with known gaze values. Then given a reference image I_a with a known gaze label $\mathbf{g}_a$ and another image I_b with unknown gaze label, their approach outputs a $\Delta\mathbf{g}_{ba}$, from which the absolute gaze for I_b can be computed as $\hat{\mathbf{y}}_b = \mathbf{y}_a + \Delta\mathbf{g}_{ba}$. Their original paper states a within-MPIIGaze error with $k = 9$ at $4.67°$ using a simple LeNet-5 style Siamese network and a pair of eye images as input. We use 256×64 eye-region images from GazeCapture as input and use a DenseNet-based architecture to make their approach more comparable to ours. Our re-implementation yields $3.53°$ for their method at $k = 9$ on MPIIGaze, a $1.2°$ improvement despite dataset differences. We show statistically significant improvements to [20] across all ranges of k in our MPIIGaze evaluations, with our method only requiring 4 calibration samples to compete with their best performance at $k = 256$ (see the red and green curves in Fig. 6). The improvement from our final approach is further emphasized in Fig. 6a with evaluations on the test subset of GazeCapture. At $k = 4$, we yield a performance improvement of 20.5% or $0.8°$ over [20].

6. Conclusion

In this paper we presented the first practical approach to deep-learning based high-accuracy personalized gaze estimation requiring only few calibration samples. Our FAZE framework consists of a disentangling encode-decoder network that learns a compact person-specific latent representation of gaze, head pose and appearance. Furthermore, we show that these latent embeddings can be used in a meta-learning context to learn a person-specific gaze estimation network from very few (as low as $k = 3$) calibration points. We experimentally showed that our approach outperforms other state-of-the-art approaches by significant margins and produces the currently lowest reported personalized gaze errors on both the GazeCapture and MPIIGaze datasets.

Acknowledgements. Seonwook Park carried out this work during his internship at Nvidia. This work was supported in part by the ERC Grant OPTINT (StG-2016-717054).

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *NeurIPS*, 2016.

[2] Margrit Betke, James Gips, and Peter Fleming. The camera mouse: visual tracking of body features to provide computer access for people with severe disabilities. *IEEE Transactions on neural systems and Rehabilitation Engineering*, 10(1):1–10, 2002.

[3] Ralf Biedert, Georg Buscher, Sven Schwarz, Jörn Hees, and Andreas Dengel. Text 2.0. In *CHI*, 2010.

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

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[6] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *ECCV*, 2018.

[7] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T. Freeman. Cognitive load estimation in the wild. In *CHI*, 2018.

[8] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017.

[9] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006.

[10] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and José MF Moura. Few-shot human motion prediction via meta-learning. In *ECCV*, 2018.

[11] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *TPAMI*, 32(3):478–500, 2010.

[12] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *ICANN*, 2011.

[13] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *CVPR*, 2018.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[15] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. Stressclick: Sensing stress from gaze-click patterns. In *ACM MM*, 2016.

[16] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *NeurIPS*, 2017.

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

[18] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[19] Kangwook Lee, Hoon Kim, and Changho Suh. Simulated+unsupervised learning with adaptive data generation and bidirectional mappings. In *ICLR*, 2018.

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

[21] Feng Lu, Takahiro Okabe, Yusuke Sugano, and Yoichi Sato. A head pose-free approach for appearance-based gaze estimation. In *BMVC*, 2011.

[22] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Inferring human gaze from appearance via adaptive linear regression. In *ICCV*, 2011.

[23] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. In *arXiv:1803.02999*, 2018.

[24] Alexandra Papoutsaki, James Laskey, and Jeff Huang. Searchgazer: Webcam eye tracking for remote studies of web search. In *CHIIR*, 2017.

[25] Eunbyung Park and Alexander C Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. In *ECCV*, 2018.

[26] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep Pictorial Gaze Estimation. In *ECCV*, 2018.

[27] 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.

[28] Anjul Patney, Joohwan Kim, Marco Salvi, Anton Kaplanyan, Chris Wyman, Nir Benty, Aaron Lefohn, and David Luebke. Perceptually-based foveated virtual reality. In *SIGGRAPH*, 2016.

[29] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A Efros, and Sergey Levine. Few-shot segmentation propagation with guided networks. *arXiv preprint arXiv:1806.07373*, 2018.

[30] Rajeev Ranjan, Shalini De Mello, and Jan Kautz. Light-weight head pose invariant gaze tracking. In *CVPR - Workshops*, 2018.

[31] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[32] Danilo Jimenez Rezende, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. One-shot generalization in deep generative models. *JMLR*, 48, 2016.

[33] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *ECCV*, 2018.

[34] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016.

[35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[36] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning

from simulated and unsupervised images through adversarial training. In *CVPR*, July 2017.

[37] John L Sibert, Mehmet Gokturk, and Robert A Lavine. The reading assistant: eye gaze triggered auditory prompting for reading remediation. In *ACM UIST*, pages 101–107, 2000.

[38] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[39] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-Synthesis for Appearance-based 3D Gaze Estimation. In *CVPR*, 2014.

[40] Kar-Han Tan, David J. Kriegman, and Narendra Ahuja. Appearance-based eye gaze estimation. In *WACV*, 2002.

[41] Ashish Tawari, Kuo Hao Chen, and Mohan M Trivedi. Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In *ITSC*, 2014.

[42] Hung-Yu Tseng, Shalini De Mello, Jonathan Tremblay, Sifei Liu, Stan Birchfield, Ming-Hsuan Yang, and Jan Kautz. Few-shot viewpoint estimation. In *BMVC*, 2019.

[43] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[44] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016.

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

[46] Michel Wedel and Rik Pieters. A review of eye-tracking research in marketing. In *Review of marketing research*. Emerald Group Publishing Limited, 2008.

[47] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Interpretable transformations with encoder-decoder networks. In *ICCV*, 2017.

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

[49] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *ETRA*, 2018.

[50] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Evaluation of appearance-based methods and implications for gaze-based applications. In *CHI*, 2019.

[51] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, 2015.

[52] 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.

[53] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *TPAMI*, 2019.

[54] Yanxia Zhang, Jörg Müller, Ming Ki Chong, Andreas Bulling, and Hans Gellersen. Gazehorizon: Enabling passers-by to interact with public displays by gaze. In *ACM UbiComp*, 2014.

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