Few-Shot Adaptive Gaze Estimation (Supplementary Material)

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

Appendix

Due to constraints on the space available in the main paper, we were unable to include all the details there. Here we provide additional implementation details pertaining to our (a) data pre-processing pipeline and (b) the configuration of our DT-ED network. We also show additional results of the ablation study (Section 5.1 in the main paper) on the test partition of the GazeCapture dataset and the performance of FAZE for the within MPIIGaze leave-one-person out setting. Finally, we show the sensitivity of FAZE to various design configurations.

A. Implementation Details

We describe further details in how we pre-process the datasets used, and the configuration of the DT-ED architecture. A reference implementation of both can be found as open-source software at https://github.com/NVlabs/few_shot_gaze.

A.1. Data Pre-processing

We employ a normalization procedure based on [12], which is a revision of [10], but with a few small changes. We utilize state-of-the-art open-source implementations for face detection[1] [4] and facial landmarks detection[2] [1], respectively. We use the Surrey Face Model [6] as the reference 3D face model, and select 4 eye corners and 9 nose landmarks as described by the Multi-PIE 68-points markup [2] for PnP-based [8] head pose estimation. This is in contrast to [10, 12] which instead use the 4 eye corners and 2 mouth corners. This is motivated by our observation that the mouth corner landmarks are not sufficiently static due to facial expression changes, and that the inherent ambiguity in determining head yaw with very few co-planar landmarks in 3D leads to less reliable head pose estimation.

In our work, we utilize a single image as input which contains both eyes. For this purpose, we select the mean of the 2 inner eye corner landmarks in 3D as the origin of our normalized camera coordinate system. We use a focal length of $1300mm$ for the normalized camera intrinsic parameters, and a distance of $600mm$ from the face to produce image patches of size 256×64 to use as input for training.

A.2. Configuration of Disentangling Transforming Encoder-Decoder (DT-ED)

We use the DenseNet architecture to parameterize our encoder-decoder network [5]. We configure the DenseNet with a growth-rate of 32, 4 dense blocks (each with 4 composite layers), and a compression factor of 1.0. We neither use dropout nor 1×1 convolutional layers. We use instance normalization [11] and leaky ReLU activation functions (with $\alpha = 0.01$) throughout the network as they proved to improve performance for all architectures.

To project CNN features back from latent features $\mathbf{z}$, we apply a fully-connected layer to output values equivalent to 32 feature maps of width 8 and height 2. The DenseNet decoder that we use to model $\mathcal{D}$ is identical in construction to a usual DenseNet but uses deconvolutional layers (with stride 1) in the place of normal convolutions, and 3×3 deconvolutions (with stride 2) instead of average pooling at the transition layers. To be faithful to the original implementation, we do not apply bias layers to convolutions in our DenseNet-based DT-ED. We initialize all layers' weights with MSRA initialization [3], while biases of the fully-connected layers are initialized with zeros.

B. Additional Results

We provide additional results of the ablation study on the test partition of the GazeCapture dataset and evaluate the within-dataset performance of FAZE on the MPIIGaze dataset.

B.1. Ablation Study on GazeCapture

In the main paper, we provide the results of the ablation study on the MPIIGaze dataset (Fig. 4 in the main paper). Our evaluation setting is a cross-dataset evaluation, where we train on the training partition of the GazeCapture dataset [7] and test on the test partition of the same dataset as well

*The first two authors contributed equally.
[1] https://github.com/cydonia999/Tiny_Faces_in_Tensorflow
[2] https://github.com/jiankangdeng/Face_Detection_Alignment

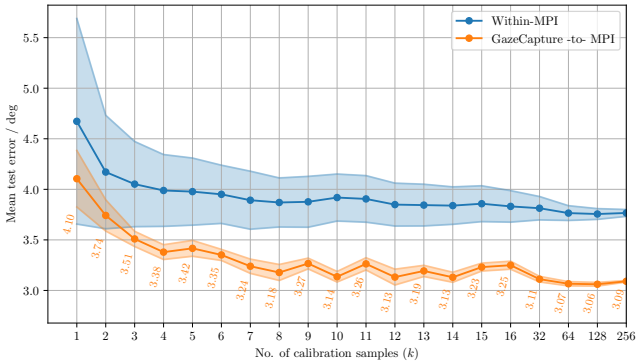


Figure 2: Gaze errors of FAZE for within-MPII leave-one-person out training (blue); and training on GazeCapture's training partition and testing one MPIIGaze (orange).

as on MPIIGaze [13]. Here we show additional results for the GazeCapture test partition (Fig. 1).

In Fig. 1a we observe the same trends for the GazeCapture test dataset that we observed for MPIIGaze. Our proposed DT-ED architecture learns latent representations that are better suited for gaze estimation than those learned by a naive encoder-decoder architecture. Additionally, for few-shot personalization significant gains in accuracy are obtained with meta-learning an adaptable network, as we propose, versus naively fine-tuning a network designed for person-independent gaze estimation (Fine-tuning versus MAML). The latter approach also leads to over-fits at very low k. Fig. 1b shows the value of our proposed loss terms of embedding consistency and of computing gaze from the latent representations while training DT-ED, for GazeCapture. Finally, Fig. 1c shows the consistent improvements obtained for the GazeCapture dataset by preserving inter-person differences versus not doing so.

B.2. Within-MPIIGaze Performance

So far Liu et al. [9] report the best known accuracy of $4.67°$ with 9 calibration samples on MPIIGaze with their differential network architecture. They use the within MPIIGaze leave-one-subject out evaluation protocol for their experiments. To directly compare against their method, we evaluate the performance of our FAZE framework for this experimental protocol (Fig. 2). With 9 calibration samples FAZE obtains a gaze error of 3.88, which is a 17% improvement over Liu et al.'s method. Note, also, that within-MPIIGaze training performs worse than training with GazeCapture (see Fig. 6 in the main paper). This is expected, given the significantly larger diversity of subjects present in the GazeCapture training subset (993) versus MPIIGaze (14 in a leave-one-out setting), which benefits both DT-ED and MAML. This observation corroborates with similar ones previously made in [7].

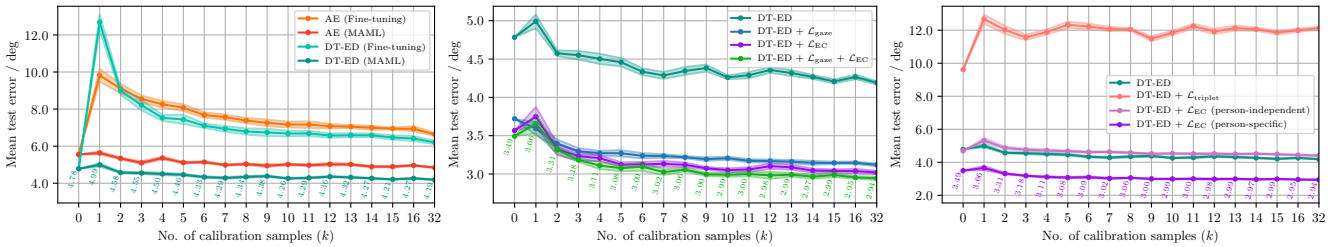
C. Sensitivity Analysis

We show the influence of various design parameters on the overall performance of our algorithm. These analyses help to determine the parameters' optimal values.

C.1. Latent Gaze Code

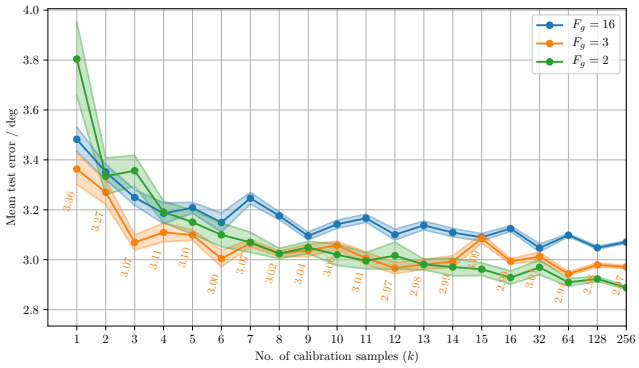
Dimension Our latent gaze code has the dimensions of $3 \times F_g$. In order to empirically select the optimal value of F_g, we evaluate the performance of FAZE for several different values of $F_g = \{16, 3, 2\}$ shown in Fig. 3, while keeping the dimensions of the appearance and head pose codes fixed at 64 and 16 respectively. Empirically we find $F_g = 2$ to be optimal for both datasets and hence select it for our final implementation.

Normalization In general we find that is beneficial to normalize our $3 \times F_g$-sized latent gaze code to achieve the lowest gaze errors. We experiment with various methods for normalization, which involve computing an ℓ_2 norm along a particular dimension and dividing all the observed values for that dimension with the norm. We compute norms along the F_g dimension resulting in 3 norms. Alternatively, one can normalize along the 3 dimension, resulting in F_g norms.

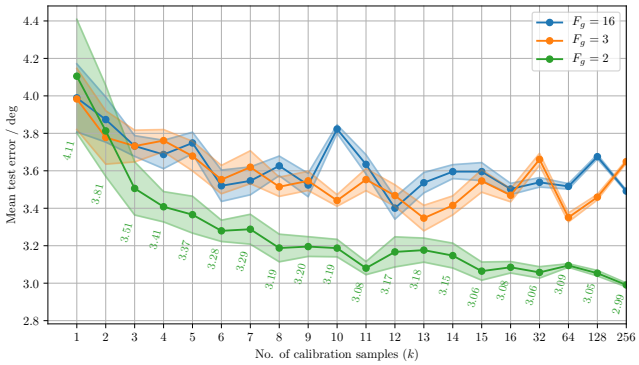


(a) (b) (c)

Figure 1: **Ablation Study on GazeCapture (test):** Impact of (a) learning the few-shot gaze estimator using MAML and using the transforming encoder-decoder for feature learning; (b) different loss terms for training the transforming encoder-decoder; and (c) comparison of the different variants of embedding consistency loss term.

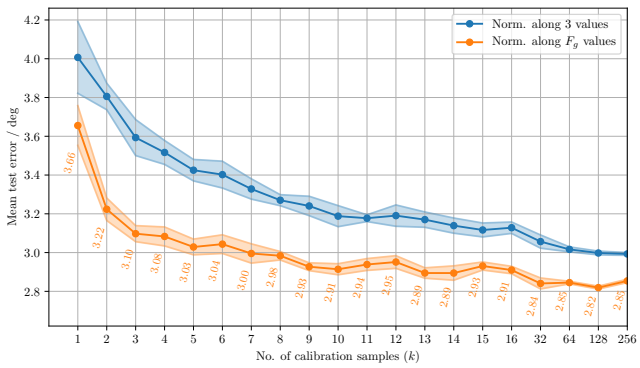


(a) GazeCapture (test)

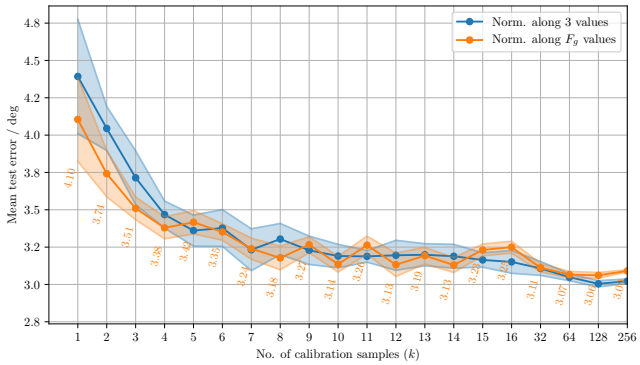


(b) MPIIGaze

Figure 3: Performance of FAZE for different dimensions F_g of the $3 \times F_g$-dimensional latent gaze code.



(a) GazeCapture (test)



(b) MPIIGaze

Figure 4: Performance of FAZE for normalizing the $3 \times F_g$-dimensional gaze code along the 3 or F_g dimensions, respectively.

We observe that normalizing along the F_g dimension, produces lower gaze errors for GazeCapture and equivalent ones for MPIIGaze, versus the alternate approach (Fig. 4). Hence, we use it for our final implementation.

References

[1] J. Deng, Y. Zhou, S. Cheng, and S. Zaferiou. Cascade multi-view hourglass model for robust 3d face alignment. In *FG*, 2018. 1

[2] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *IVC*, 28(5):807–813, May 2010. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 1

[4] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *CVPR*, 2017. 1

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 1

[6] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In *VISIGRAPP*, 2016. 1

[7] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye Tracking for Everyone. In *CVPR*, 2016. 1, 2

[8] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009. 1

[9] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation with calibration. In *BMVC*, 2018. 2

[10] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-Synthesis for Appearance-based 3D Gaze Estimation. In *CVPR*, 2014. 1

[11] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1

[12] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *ETRA*, 2018. 1

[13] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, 2015. 2