Supplementary Material: Learning-by-Novel-View-Synthesis for Full-Face Appearance-Based 3D Gaze Estimation

1. Ablation Study with the Mask Model

Ablations\Datasets	MPII-NV	XGazeF-NV
Black	16.8 (↓ 35%)	12.0 (↓ 44%)
+ Color (1:1)	14.2 (↓ 20%)	11.5 (↓ 39%)
+ Scene (1:1:3)	12.9 (↓ 10%)	9.0 (↓ 30%)
+ Weak-light	12.7 (↓ 9%)	8.3 (↓ 26%)

Table 1. Ablation study for analyzing the data augmentation. The data augmentation components are evaluated using the mask-guided model. The percentages indicate the error reduction from the baseline network under the same conditions.

In the main paper, we showed an ablation study using the baseline gaze estimation model trained on MPII-NV and XGazeF-NV. We further show the ablation study results using the proposed mask-guided model. Table. 1 shows the gaze estimation errors trained with MPII-NV (tested on the ETH-XGaze Train set) and XGazeF-NV (tested on the ETH-XGaze Test set). The percentages indicate the error reduction from the baseline network under the same conditions. Compared with the baseline network performance, it can be seen that the mask-guided model is more effective when there is only black-background training data. Although the error reduction from the baseline gets smaller, each data augmentation consistently reduces this error. This proves again that the proposed method takes full benefit from synthetic data together with the data augmentation.

2. Analysis of the EYEDIAP FT Dataset

As discussed in the paper, one of the limitations is that our synthetic dataset did not outperform other real datasets when tested on the EYEDIAP FT dataset. EYEDIAP FT employs a floating physical gaze target and has large offsets between head pose and gaze. *I.e.*, participants tend to direct their eyes to follow floating targets instead of heads in EYEDIAP FT. As a visual explanation, Fig. 1 shows the offset distributions of MPIIFaceGaze [4], EYEDIAP FT, and ETH-XGaze. Since the offset distribution is expected to be independent of the camera position, we visualize the ETH-XGaze distribution using the frontal camera. It can be seen that EYEDIAP FT has a wider offset range compared

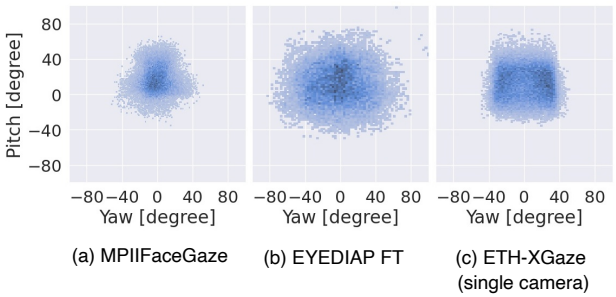

Figure 1. The offset distributions of (a) MPIIFaceGaze, (b) EYEDIAP FT, and (d) ETH-XGaze (single camera).

Training \Test	ETH-XGaze (Test)
MPII-NV (EYEDIAP FT)	**17.5**
MPII-NV ($\sigma = 5$)	20.8
MPII-NV ($\sigma = 10$)	18.4
MPII-NV ($\sigma = 20$)	20.3
MPII-NV ($\sigma = 30$)	20.3
MPII-NV ($\sigma = 40$)	22.7

Table 2. Gaze estimation errors in degree tested on EYEDIAP FT using the Baseline model.

to MPIIFaceGaze using screen-based gaze target. As discussed in the paper, therefore, the target gaze range cannot be fully covered by synthesizing samples only according to the head pose distribution. In contrast, since ETH-XGaze is also collected using a screen-based target, its offset is wider but still in a feasible range compared to MPII-NV.

Table. 2 further shows the performance of the baseline models trained on the datasets sampled from a normal distribution as Section 5.3 in the paper. While the performance naturally degrades with the small $\sigma = 5$, it does not improve with the large $\sigma = 40$, either. Considering the difference from ETH-XGaze, this indicates that the fundamental issue is the lack of facial appearance corresponding to extreme gaze directions. Although our approach can cover the gaze range of EYEDIAP FT with large σ, this results in many extreme head poses which do not appear in the original EYEDIAP FT. This facial appearance gap is likely to lead to performance degradation.

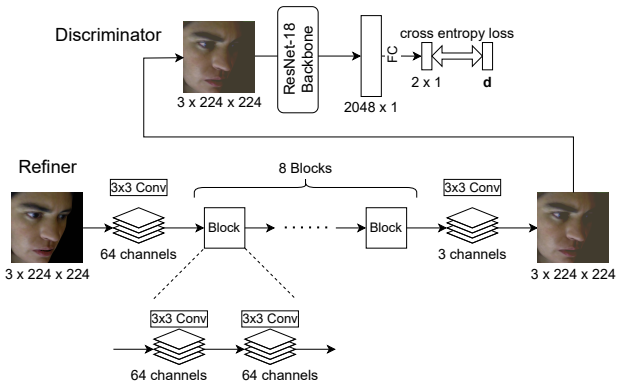

Figure 2. Architecture of our SimGAN baseline. The discriminator is based on ResNet-18 architecture, and the refiner consists of a stacked convolution blocks.

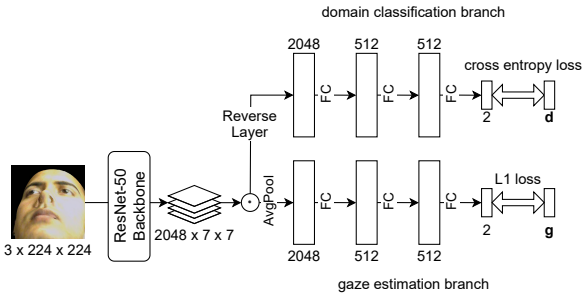

Figure 3. Architecture of our DANN baseline. We take the features before the last FC layer of the ResNet-50 backbone, and use three FC layers for domain classification and gaze estimation.

3. Baseline Implementation Details

SimGAN [3] Fig. 2 shows the architecture of our SimGAN implementation. The discriminator is based on ResNet-18 architecture, and the refiner is a stacked convolution blocks whose total depth is the same as ResNet-18. To train the refiner, we use a loss function $\mathcal{L} = \mathcal{L}_a + \gamma\mathcal{L}_i$, where $\mathcal{L}_a$ is adversarial loss, and $\mathcal{L}_i$ is an $\ell 1$ loss defined between the source and refined images. We empirically set the loss weight $\gamma = 8.0$.

DANN [1] Fig. 3 shows the architecture of our DANN implementation. We used the ResNet-50 backbone and three FC layers for both domain classification and gaze estimation. We also used batch normalization and ReLU activation layers between FC layers. The reverse layer indicates the operation to multiply -1 with the gradient of domain classification loss, which is a cross entropy loss between predicted domain label and domain label $\boldsymbol{d}$. Therefore, the feature extraction network is trained in an adversarial way to the domain classification loss. To train this model, we use a loss function $\mathcal{L} = \mathcal{L}_g + \gamma\mathcal{L}_d$, where $\mathcal{L}_g$ is an $\ell 1$ loss for gaze estimation, and $\mathcal{L}_d$ is a cross entropy loss for domain classification. We empirically set the loss weight $\gamma = 1.0$.

PADACO [2] Our PADACO implementation has the same network architecture as DANN but a different sampling method for the source dataset. The network is first pre-trained with gaze estimation loss on the source dataset and outputs predictions from target samples. We then calculate the sampling probabilities of each source data based on the gaze prediction results on the target data. We count the 10 nearest source samples for each target prediction and define the sampling probability according to the number of times each source sample appears in the neighborhood. Finally, the whole network is trained with both source and target data, where the source data is sampled according to the sampling probabilities. We use the same loss function as DANN, and empirically set the loss weight $\gamma = 0.9$.

References

[1] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. ICML*, 2015. 2

[2] 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

[3] 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. 2

[4] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE TPAMI*, 41(1):162–175, 2019. 1

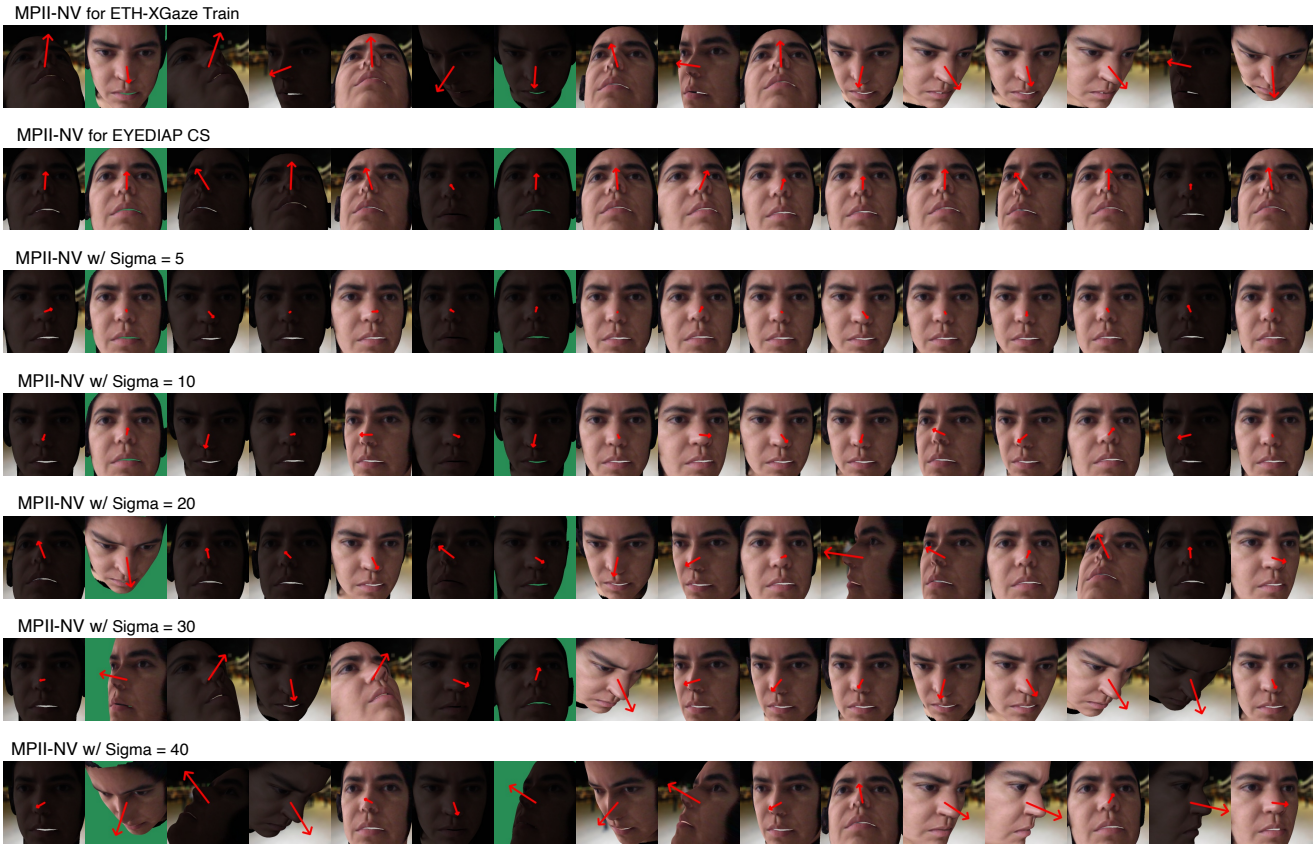

Figure 4. Same random augmentation for the compared datasets.