# Supplementary Materials: Rotation-Constrained Cross-View Feature Fusion for Multi-View Appearance-based Gaze Estimation

Yoichiro Hisadome, Tianyi Wu, Jiawei Qin, Yusuke Sugano
Institute of Industrial Science, The University of Tokyo
Komaba 4-6-1, Tokyo, Japan
{hisadome, twu223, jqin, sugano}@iis.u-tokyo.ac.jp

| Head pose | Seen | Unseen |
|---|---|---|
| w/o Separate Fusers | 3.49° | 5.10° |
| w/o Backbone Features | **3.46°** | 5.20° |
| Proposed | 3.50° | **4.95°** |

Table 1. Ablation studies of learnable modules. We ablated the separate weight of the *Fusers* and *3D Feature Extractor*.

## 1. Detailed Ablation Studies

We perform ablation studies on several learnable modules of the proposed method to validate our design choice on *Fuser*. The first row (*w/o Separate Fusers*) in Table 1 corresponds to a variant of the proposed method where the *Fusers* in each block share the same weights. The model in the second row (*w/o Backbone Features*) uses the initial rotatable feature $\mathbf{F}^{(0)}$ as input to *Fusers* instead of the backbone feature $\mathbf{f}$. This model, therefore, does not distinguish between rotatable and backbone features.

While both methods perform on par with the proposed method under the *seen* setting, the proposed method shows superiority in the *unseen* setting. One possible explanation is that stacking different fusion blocks allows the model to focus on different patterns depending on the depth of the block and that the original backbone feature still contains valuable information for appearance-based gaze estimation.

## 2. Visualization of Rotatable Features

In Fig. 1, we depict more Isomap embedding of the initial rotatable features from test subjects. Each Isomap embedding was generated from the features obtained from each target participant, and all other visualization details are consistent with the main paper. The visualization results confirm that the proposed method acquires person-independent rotatable feature representations.

In Fig. 2, we also show more scatter plot visualizations of the rotatable features from test subjects in the yaw-pitch coordinate system. We can consistently observe the tendency for feature distributions to converge before the first fusion block and then diverge in later blocks across different subjects. It can be seen that the proposed method dynamically updates rotatable features even with a slight rotation (the upper right example in Fig. 2).

## 3. Baseline Implementation Details

Unless otherwise noted, all baseline methods follow the same training hyperparameters as used for the proposed method in the main paper. We note that we did not tune the hyperparameters in favor of the proposed method. Instead, we used common choices, most of which already comply with ResNet and PureGaze. With the Cyclic LR scheduler, ResNet, PureGaze, and Hybrid-TR are less tuning demanding. Therefore we tune the training-unstable DT-ED.

**DT-ED** Since we use a richer full-face patch instead of an eye-region patch as input of DT-ED, we modified the appearance and gaze latent code sizes from 64 and 2 to 512 and 16. Following the original setting, we used angular loss for gaze estimation and $\ell_1$ for reconstruction. For the learning rate, we found that the scaling and ramp-up settings in the original paper make it difficult for the model to reconstruct the target image. Therefore, we trained the model with a base learning rate of $5 \times 10^{-4}$ decaying by 0.8 every 1 epoch, similar to another gaze redirection work [4]. Unlike other baselines, the batch size is set to 60.

**Gaze-TR** In our implementation, we used ResNet-50 [1] to extract feature maps from the images. The size of the feature map was $7 \times 7 \times 32$, which is then fed to a six-layer transformer. Finally, an MLP takes the feature vector as input and estimates the gaze direction.

**PureGaze** When training models on both ETH-XGaze [3] and MPII-NV [2] dataset, we used the

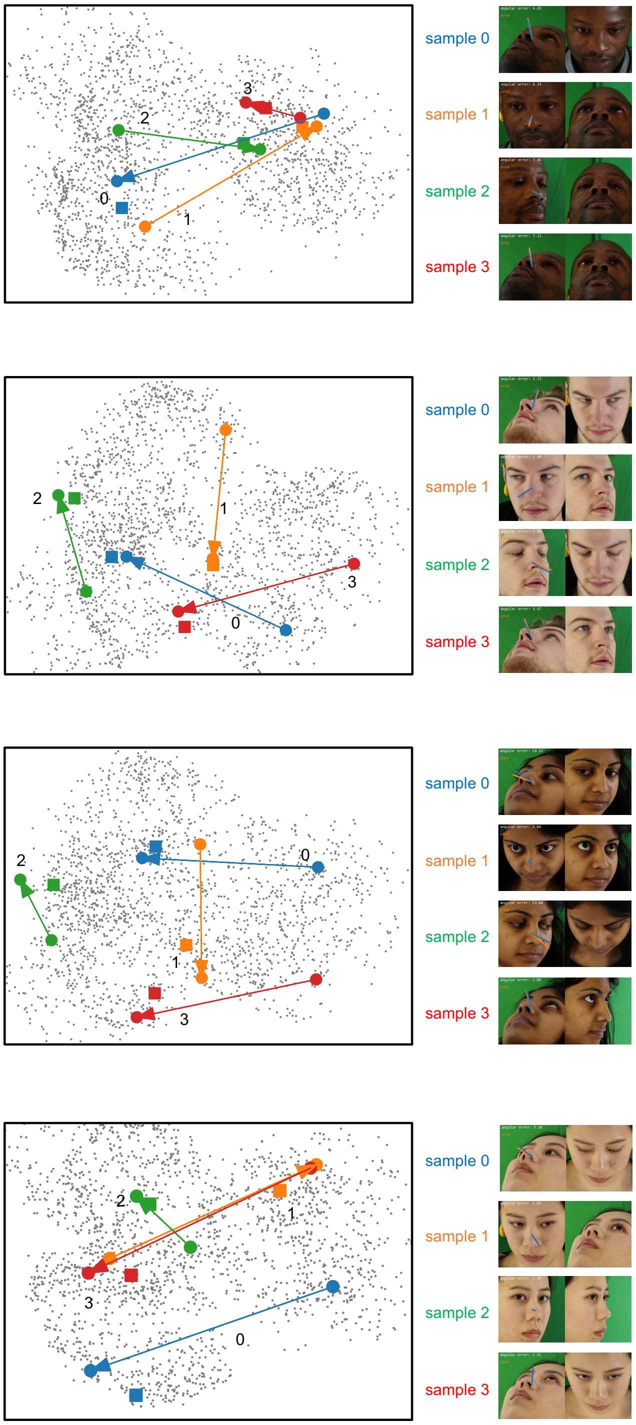

Figure 1. Isomap embedding of the initial rotatable features. The right side shows the example input samples. $\mathbf{F}^{(0)}_{\text{ref}}$, $\mathbf{F}^{(0)}_{\text{tgt}}$, and $\mathbf{RF}^{(0)}_{\text{ref}}$ of the same sample are represented in the same color on the left side plot.

default mask image in the official PureGaze repository [1] generated for normalized ETH-XGaze face images to compute the adversarial reconstruction loss. For the extra hyperparameters controlling the relative contribution of the adversarial loss to the total loss, we followed the official implementation.

## 4. Definition of the Rotation Matrix

As discussed in the paper, there are two approaches to computing the relative rotation matrix $\mathbf{R}$ using either camera calibration or head poses estimation. In the following, we provide detailed explanations of two claims: 1) the final $\mathbf{R}$ becomes the same in either approach, and 2) the relative translation $\mathbf{t}$ is uniquely determined by $\mathbf{R}$ and can be ignored.

First, we show that the two definitions $\mathbf{R} = \mathbf{N}_{\text{tgt}}\tilde{\mathbf{R}}\mathbf{N}^{\top}_{\text{ref}}$ and $\mathbf{R} = \mathbf{H}_{\text{tgt}}\mathbf{H}^{\top}_{\text{ref}}$ are interconvertible. Let us denote the camera extrinsic parameters, *i.e.*, the transformation from the reference to the target camera coordinate systems, as $\mathbf{C} \in \mathbb{R}^{4\times4}$. If we denote head poses in the original camera coordinate systems before normalization as $\hat{\mathbf{H}} \in \mathbb{R}^{4\times4}$, their relationship can be defined as

$$\hat{\mathbf{H}}_{\text{tgt}} = \mathbf{C}\hat{\mathbf{H}}_{\text{ref}}. \tag{1}$$

If we further denote an extended $4 \times 4$ normalization matrix as $\mathbf{N}$, the head poses $\mathbf{H}$ after normalization can also be obtained from the normalization matrix as

$$\mathbf{H} = \mathbf{N}\hat{\mathbf{H}}. \tag{2}$$

From Eq. 1 and Eq. 2, we can derive that

$$\begin{aligned} \mathbf{N}_{\text{tgt}}\mathbf{C}\mathbf{N}^{\top}_{\text{ref}} &= \mathbf{N}_{\text{tgt}}\hat{\mathbf{H}}_{\text{tgt}}\hat{\mathbf{H}}^{\top}_{\text{ref}}\mathbf{N}^{\top}_{\text{ref}} \\ &= (\mathbf{N}_{\text{tgt}}\hat{\mathbf{H}}_{\text{tgt}})(\mathbf{N}_{\text{ref}}\hat{\mathbf{H}}_{\text{ref}})^{\top} \\ &= \mathbf{H}_{\text{tgt}}\mathbf{H}^{\top}_{\text{ref}}. \end{aligned}$$

Therefore, we can conclude that the two definitions are interconvertible and have the same meaning. Note that this applies not only to the rotation component $\mathbf{R}$ but also to the translation component $\mathbf{t}$.

Next, we show that the translation component $\mathbf{t}$ is uniquely determined by the rotation $\mathbf{R}$ under the assumption of data normalization. One of the key properties of the normalization process is that the origin of the gaze vector is located at a fixed distance $d$ on the $z$-axis of the camera coordinate system. Therefore, this origin $\mathbf{o} = (0, 0, d, 1)^{\top}$ does not move when the above transformation matrix is applied:

$$\mathbf{o}_{\text{tgt}} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{pmatrix} \mathbf{o}_{\text{ref}} = \mathbf{o}_{\text{ref}}. \tag{3}$$

If we denote $\mathbf{R} = (\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z)$ where $\mathbf{r}_x, \mathbf{r}_y, \mathbf{r}_z \in \mathbb{R}^3$ are the column vectors of the rotation matrix, substituting this

[1] https://github.com/yihuacheng/PureGaze

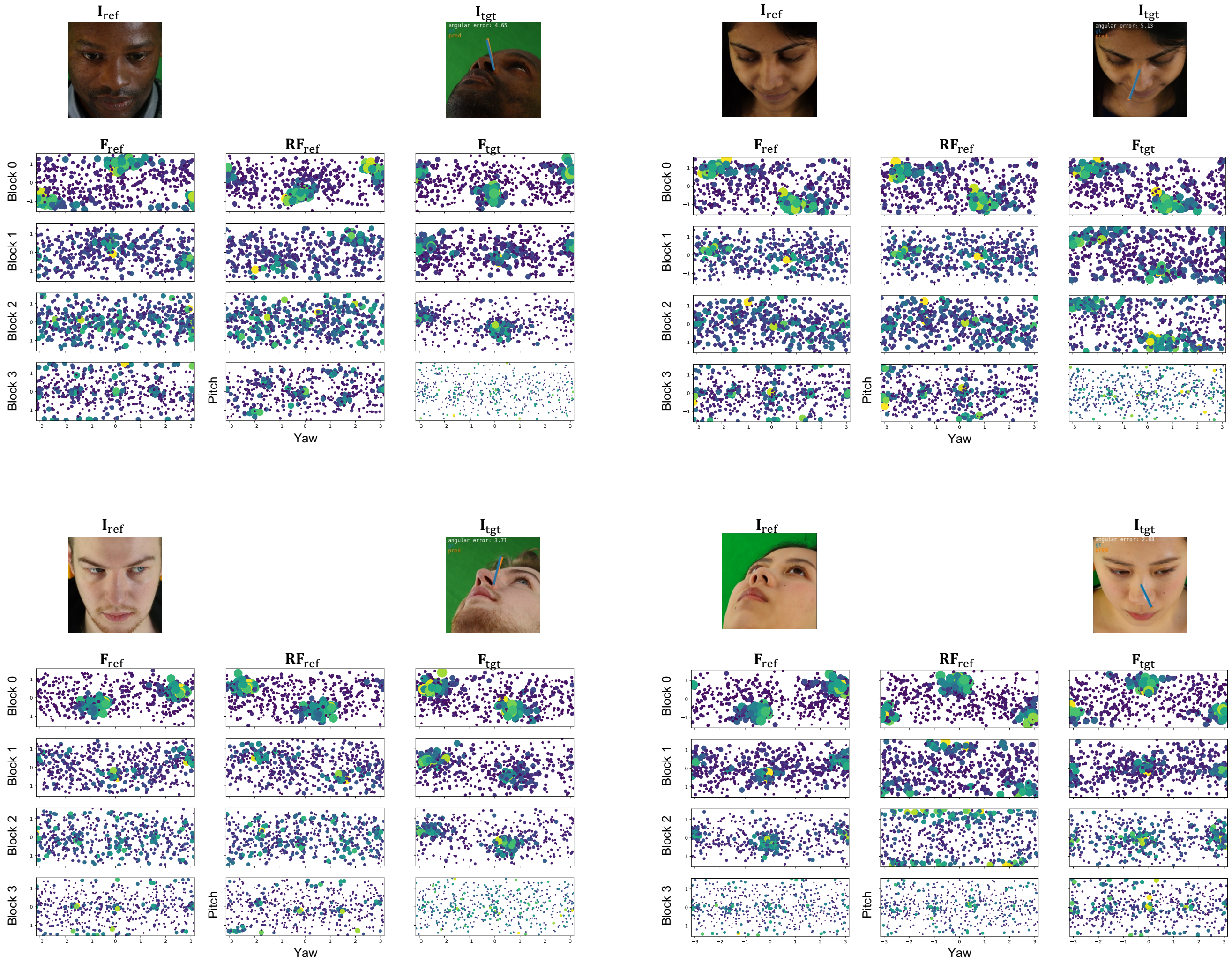

Figure 2. Scatter plot visualization of the rotatable features. Each of the $D$ 3D vectors is represented in a pitch-yaw coordinate system. Each row corresponds to the rotatable features at different fusion stages. Larger and yellower dots represent elements with a larger norm.

into Eq. 3 yields

$$\begin{aligned}\begin{pmatrix}0\\0\\d\\1\end{pmatrix} &= \begin{pmatrix}\mathbf{r}_x & \mathbf{r}_y & \mathbf{r}_z & \mathbf{t}\\ & 0 & & 1\end{pmatrix}\begin{pmatrix}0\\0\\d\\1\end{pmatrix}\\ &= \begin{pmatrix}d\mathbf{r}_z+\mathbf{t}\\ 1\end{pmatrix}.\end{aligned}$$

Therefore, the translation component $\mathbf{t}$ is uniquely defined by the fixed distance $d$ and the rotation vector $\mathbf{r}_z$ as

$$\mathbf{t} = \begin{pmatrix}0\\0\\d\end{pmatrix} - d\mathbf{r}_z, \tag{4}$$

and can be ignored in our problem setting.

## References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 1

[2] Jiawei Qin, Takuru Shimoyama, and Yusuke Sugano. Learning-by-novel-view-synthesis for full-face appearance-based 3d gaze estimation. In *Proc. CVPRW*, 2022. 1

[3] 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

[4] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. In *Proc. NIPS*, 2020. 1