Fine Gaze Redirection Learning with Gaze Hardness-aware Transformation (Appendix)

1. Redirection Process

We use the transforming autoencoder (TA) structure proposed by Hinton *et al.* [2] as a backbone for gaze redirection. Similar to FAZE and STED [6, 15], the redirection process (R in main body) of TA responsible for the transformation of latent features is defined based on the rotation matrix $\boldsymbol{R}$ in Eq. 1.

$$\tilde{\mathbf{z}}_t^g = \mathrm{R}(\mathbf{z}_s^g) = \boldsymbol{R}_t^g \left(\boldsymbol{R}_s^g\right)^{-1} \mathbf{z}_s^g, \tag{1}$$

where the rotation matrix $\boldsymbol{R}_s^g$ for the transformation of the source gaze feature $\mathbf{z}_s^g$ is as follows:

$$\boldsymbol{R}_s^g = \begin{pmatrix} \cos\phi_s^g & 0 & \sin\phi_s^g \\ 0 & 1 & 0 \\ -\sin\phi_s^g & 0 & \cos\phi_s^g \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_s^g & -\sin\theta_s^g \\ 0 & \sin\theta_s^g & \cos\theta_s^g \end{pmatrix} \tag{2}$$

Here, (θ_s^g, ϕ_s^g) represents the pitch and yaw angle of the source gaze direction. Similar to Eq. 2, $\boldsymbol{R}_t^g$ is defined based on (θ_t^g, ϕ_t^g). The redirection process based on these rotation matrices is also applied to head pose $\mathbf{z}_s^h$. Redirected features are used to generate redirected images for supervised learning.

2. Other loss function

This section describes $\mathcal{L}_{other}$ for better reconstruction. $\mathcal{L}_{other}$ includes pixel-wise reconstruction loss and perceptual loss, which is defined as follows [15]:

$$\mathcal{L}_{other} = \|\widetilde{\mathbf{x}}_t - \mathbf{x}_t\|_1 + \sum_{k=1}^{5} \|F_k(\widetilde{\mathbf{x}}_t) - F_k(\mathbf{x}_t)\|_2 \tag{3}$$

where $F_i(\cdot)$ is the activation feature map of the i-th layer of ψ.

3. Further Analysis of SG loss

Relationship with contrastive loss. Let's analyze the operation of $\mathcal{L}_{sg}$ in terms of the well-known contrastive loss. First, $J_{i,j}$ of $\mathcal{L}_{sg}$ is rearranged using logarithmic and exponential operators as follows:

$$\begin{aligned} J_{i,j} &= D_{i,j} + \sum_{(i,k)\in\mathcal{N}} \log(e^{\delta - D_{i,k}}) + \sum_{(j,l)\in\mathcal{N}} \log(e^{\delta - D_{j,l}}) \\ &\leq D_{i,j} + \log \sum_{(i,k)\in\mathcal{N}} e^{\delta - D_{i,k}} + \log \sum_{(j,l)\in\mathcal{N}} e^{\delta - D_{j,l}} \end{aligned} \tag{4}$$

Note that temperature hyper-parameter τ was omitted for simplicity. In Eq. 4, only the cases where $D_{i,k}$ and $D_{j,l}$ are always less than a margin δ are considered admissible pairs. Meanwhile, the second and third terms (terms with negative pairs) are changed to $\mathrm{LogSumExp}$ form with a tight upper-bound range. Therefore, $\mathcal{L}_{sg}$ can be redefined by

$$\begin{aligned} \mathcal{L}_{sg} =& \frac{1}{2|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} D_{i,j} + \frac{1}{2|\mathcal{P}|} \sum_i \log \sum_k e^{\delta - D_{i,k}} \\ &+ \frac{1}{2|\mathcal{P}|} \sum_j \log \sum_l e^{\delta - D_{j,l}} \end{aligned} \tag{5}$$

Eq. 5 only deals with the case where $J_{i,j} > 0$, and hard negatives are mined by $\max(0, J_{i,j})$. The first term in Eq. 5 is learned so that $D_{i,j}$ is minimized. On the other hand, since $e^{\delta - D_{i,k}}$ and $e^{\delta - D_{j,l}}$ of the second and third terms are minimized, each of $D_{i,k}$ and $D_{j,l}$ is learned toward the increasing direction [4].

Note that Eq. 5 is considered as a form of *generalized* contrastive loss with mined hard negatives. In detail, the first term of RHS in Eq. 5 is an alignment term that encourages the gaze direction of positive pairs $(\mathbf{z}_s^g, \mathbf{z}_s^+)$ or $(\mathbf{z}_s^g, \mathbf{z}_s^e)$ to be consistent (see Figure 3(b) in the main body). The second and third terms are regarded as distribution matching terms that encourage the distribution of negative pairs to match the prior distribution [10]. In particular, terms with $\mathrm{LogSumExp}$ encourage latent feature representations to match uniform distributions on the hypersphere. As a result, the second and third terms of Eq. 5 are trained so that the distribution of negative pairs $(\mathbf{z}_s^g, \mathbf{z}_s^h)$, $(\mathbf{z}_s^g, \mathbf{z}_s^u)$ and $(\mathbf{z}_s^g, \mathbf{z}_s^-)$ matches the uniform distribution. Therefore, $\mathbf{z}_s^g$ enables to utilize unbiased hard negatives for similarity learning.

Analysis from an information-theoretic perspective. It is known that contrastive loss has a lower bound of mutual information [5, 7]. Similarly, the generalized contrastive loss

Table 1: Performance according to the mini-batch size in the test split of GazeCapture dataset

Mini-batch	err_g	err_h	$h \to g$	$g \to h$	LPIPS
32	1.973	0.720	1.933	0.334	**0.196**
48	2.010	0.770	1.895	0.380	0.203
64	2.110	0.724	1.993	0.340	0.206
128	**1.964**	**0.693**	**1.882**	**0.330**	0.203

in Eq. 5 can also be interpreted as the mutual information (I) with entropy (H) between two latent variables, i.e., U and V: $I(U, V) = H(U) - H(U|V)$. The alignment term in Eq. 5 is directly related to $H(U|V)$, which aims to reduce the uncertainty between positive pairs. Distribution matching terms are related to $H(U)$ and can be considered auxiliary pairs to maximize entropy. Therefore, Eq. 5 has a (compact) lower bound based on mutual information and theoretically guarantees learning stability.

4. Implementation of ContraCAM

We employed a class activation map (CAM) to visualize the effect of discriminative learning of the proposed method. Unlike CAM [16] and Grad-GAM [8], which use the discrete probability of Softmax as a confidence score, ContraCAM [3] uses continuous probability as a confidence score. So, ContraCAM is suitable for visualizing the activation map of the proposed method because it can utilize gaze and head pose predictions as continuous confidence scores. ContraCAM is defined by

$$\mathrm{ContraCAM}_{hw} = \mathrm{Normalize}\left(\mathrm{ReLU}\left(\sum_c \alpha_c A^c_{hw}\right)\right)$$
$$\alpha_c = \mathrm{ReLU}\left(\frac{1}{HW}\sum_{h,w}\frac{\partial \mathrm{MLP}(\mathbf{z}_s^{g/h/u})}{\partial A^c_{hw}}\right) \tag{6}$$

where A^c_{hw} is the feature map or spatial activation extracted from the middle stage (the 6th layer) of encoder $\mathcal{E}$. Also, h, w and c indicate the index of height (H), width (W) and channel size (C), respectively. $\mathrm{Normalize}(x) = \frac{x - \min(x)}{\max(x) - \min(x)}$ is a normalization function that maps the range of x to $[0, 1]$. $\mathrm{MLP}(\cdot)$ stands for multi-layer perceptron (MLP) that extracts a confidence score (prediction) from each feature. In this paper, the number of layers of MLP is 2. The main differences between ContraCAM and original CAM in Eq. 6 are as follows: One is that clips activation values with a non-negative sign and another is $\mathrm{MLP}(\cdot)$ (marked as red color) that outputs (continuous) confidence scores. We can see the Pytorch-like pseudocode

Table 2: Performance according to metric loss

Metric loss	err_g	err_h	$h \to g$	$g \to h$	LPIPS
Margin [11]	2.264	0.827	1.994	0.368	0.212
DSML (tri) [13]	2.100	0.799	1.915	0.377	0.206
SG (Ours)	**1.973**	**0.720**	**1.933**	**0.334**	**0.196**

that describes the behavior of ContraCAM in Listing 1.

Figure 3 shows additional qualitative results using ContraCAM on the GazeCapture dataset. In all samples, gaze features of the proposed method gave higher attention scores to the eye region than STED. We can observe that while the head pose and task-irrelevant features of STED consider the eye region together, the features of the proposed method separate the eye region and other regions from each other.

```
A = encoder[:-2](x)  # 6th feature map of encoder

z = encoder(x)  # feature vector

predicted_label = MLP(z)

grad = autograd.grad(predicted_label.sum(), A)[0]

weight = adaptive_avg_pool2d(grad, output_size
    =(1, 1))
weight = relu(weight)  # non-negative clipping

# weighted sum
cam = sum(weight * A, dim=1, keepdim=True).detach
    ()

cam = resize(cam(h, w))
cam = normalize(relu(cam)) # normalize to [0, 1]
```

Listing 1: Pytorch-style pseudo-code for ContraCAM

5. Additional Experiments

5.1. Abalation Study

Variation of mini-batch size. Table 1 shows the performance change as the mini-batch size increases in the GazeCapture dataset. As the mini-batch size increased from 32 to 128, the overall performance of all metrics improved. In the case of 48 and 64, the performance changed marginally, but when 128 was used, we could achieve an average 6% performance improvement in almost all metrics compared to 32.

Other metric losses. In order to verify the discriminative learning ability of the proposed SG loss, gaze redirection was performed through different metric losses (see Table 2). First, the formula for margin loss [11] is as follows:

$$\mathcal{L}_{margin} = \left[\alpha + y_{i,j}\left(D_{i,j} - \beta\right)\right]_+ . \tag{7}$$

where the flexible boundary parameter β is learnable, and the static margin α is fixed to 1. Positive and negative class indicator is $y_{i,j} \in \{-1, 1\}$. Next, deep SNR-based metric learning (DSML) of [13] measures the similarity between two features using the SNR metric $d_S(\mathbf{z}_i, \mathbf{z}_j) = \frac{var(\mathbf{z}_i - \mathbf{z}_j)}{var(\mathbf{z}_i)}$ rather than the Euclidean distance. Here, $var(\mathbf{z})$ is the variance of $\mathbf{z}$. The triplet-based SNR metric loss we adopted is as follows:

$$\mathcal{L}_{DMSL(tri)} = \left[d_S(\mathbf{z}_s^g, \mathbf{z}_s^e) - d_S(\mathbf{z}_s^g, \mathbf{z}_s^h) + \alpha\right]_+ + \left[d_S(\mathbf{z}_s^g, \mathbf{z}_s^e) - d_S(\mathbf{z}_s^g, \mathbf{z}_s^u) + \alpha\right]_+ . \quad (8)$$

where margin α was set to 1. As in Table 2, the proposed SG loss achieved about 12% lower err_g than the margin loss. In addition, when using triplet-based DSML, an average 8% improvement in performance was observed in all metrics.

Finally, we evaluated the performance according to the use of M^h and M^u in Eq. 3 of main body. When M^h and M^u was used, err_g was 1.884, which is about 10.3% higher than 2.101 when not used. When $\mathbf{z}_s^e$ was used instead of $\mathbf{z}_s^g$ in Eq. 3 of main body, there was a slight performance difference of about 0.97%.

5.2. Within-dataset Evaluation

We compared the performance of the proposed method and state-of-the-art redirection methods according to the within-dtaset evaluation protocol. Table 3 shows the performance of the proposed method and other methods on the MPI-IGaze, Columbia and EYEDIAP datasets. Note that the proposed method showed consistently better performance for all datasets. Therefore, not only Table 3, but also the cross-dataset evaluation result of the main body, which is a more difficult evaluation, sufficiently proves the outstanding performance of the proposed method.

5.3. Interpolation and Extrapolation

We use the interpolated gaze feature $\mathbf{z}_{tr}^g$ to generate a redirected image $\widetilde{\mathbf{x}}$: $\widetilde{\mathbf{x}} = \mathcal{G}(\mathrm{Concat}(\mathbf{z}_{tr}^g, \mathbf{z}_s^h, \mathbf{z}_s^u))$. The results of Figure 1(a) suggest that the proposed method manipulates the gaze direction between the source and the target well while maintaining the identity of the generated face. Results for more samples are given in Figure 4. Also, Figure 1(b) shows the image generated using the extrapolated gaze feature between the source and the target. We can observe that the proposed method can generate images with gaze direction that are not limited to source and target images.

5.4. Gaze Direction of Generated Gaze Feature

To prove the reliability of the gaze direction of the image generated using the interpolated gaze feature, we calculated the difference between the GT and the gaze direction of the image generated by the pre-trained gaze estimation network ψ. This experiment used gaze features generated from 1000 source and target image pairs randomly sampled from the test set of the GazeCapture dataset. Figure 2 plots the strong correlation between the predicted gaze direction and GT (Pearson correlation coefficient of 0.93). This proves that the interpolated gaze feature represents the corresponding gaze direction well.

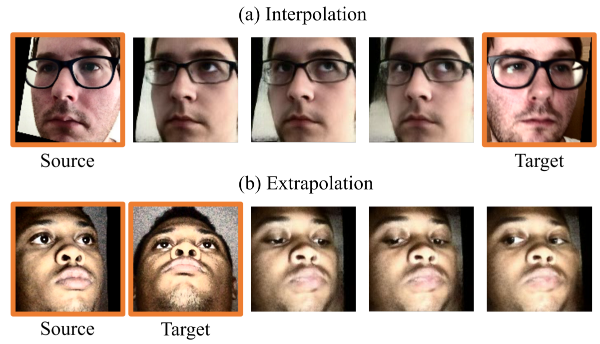

Figure 1: Generated image using interpolated and extrapolated gaze feature.

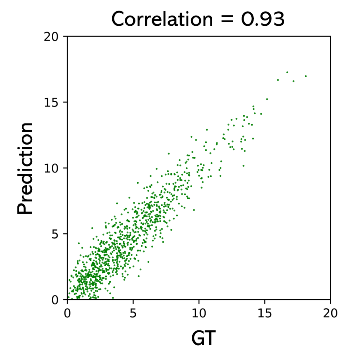

Figure 2: Correlation with ground-truth and predicted gaze direction from the image generated using interpolated gaze feature.

References

[1] 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 *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258, 2014.

[2] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International conference on artificial neural networks*, pages 44–51. Springer, 2011.

[3] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. *arXiv preprint arXiv:2108.00049*, 2021.

[4] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

Table 3: Quantitative results of within-dataset evaluation protocol on MPIIGaze, Columbia and EYEDIAP dataset. The percentage indicates the degree of improvement of the proposed method compared to STED.

Method	err_g	$u \to g$	$h \to g$	err_h	$u \to h$	$g \to h$	LPIPS
GazeFlow [12]	5.887	3.778	5.312	3.713	1.714	3.121	0.243
FAZE [6]	7.312	-	6.714	2.512	-	1.917	0.237
STED [15]	2.133	0.605	2.312	0.724	0.314	0.442	0.204
Ours	**1.814**	**0.512**	**1.994**	**0.684**	**0.211**	**0.339**	**0.202**

(a) MPIIGaze [14]

Method	err_g	$u \to g$	$h \to g$	err_h	$u \to h$	$g \to h$	LPIPS
GazeFlow† [12]	7.312	-	-	5.076	-	-	0.274
FAZE [6]	6.914	-	4.814	3.114	-	2.997	0.247
STED [15]	3.134	0.902	3.307	**0.886**	0.334	1.002	0.233
Ours	**2.872**	**0.782**	**2.902**	0.902	**0.314**	**0.987**	**0.212**

(b) Columbia [9]

Method	err_g	$u \to g$	$h \to g$	err_h	$u \to h$	$g \to h$	LPIPS
GazeFlow† [12]	17.12	-	-	3.124	-	-	0.264
FAZE [6]	16.985	-	15.625	2.962	-	2.493	0.239
STED [15]	13.094	6.413	12.796	0.817	0.662	1.674	**0.224**
Ours	**11.094**	**5.498**	**9.438**	**0.802**	**0.403**	**0.904**	0.232

(c) EYEDIAP [1]

[5] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[6] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019.

[7] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[9] Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280, 2013.

[10] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[11] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017.

[12] Yong Wu, Hanbang Liang, Xianxu Hou, and Linlin Shen. Gazeflow: Gaze redirection with normalizing flows. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[13] Tongtong Yuan, Weihong Deng, Jian Tang, Yinan Tang, and Binghui Chen. Signal-to-noise ratio: A robust distance metric for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4815–4824, 2019.

[14] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.

[15] Yufeng Zheng, Seonwook Park, Xucong Zhang, Shalini De Mello, and Otmar Hilliges. Self-learning transformations for improving gaze and head redirection. *Advances in Neural Information Processing Systems*, 33, 2020.

[16] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

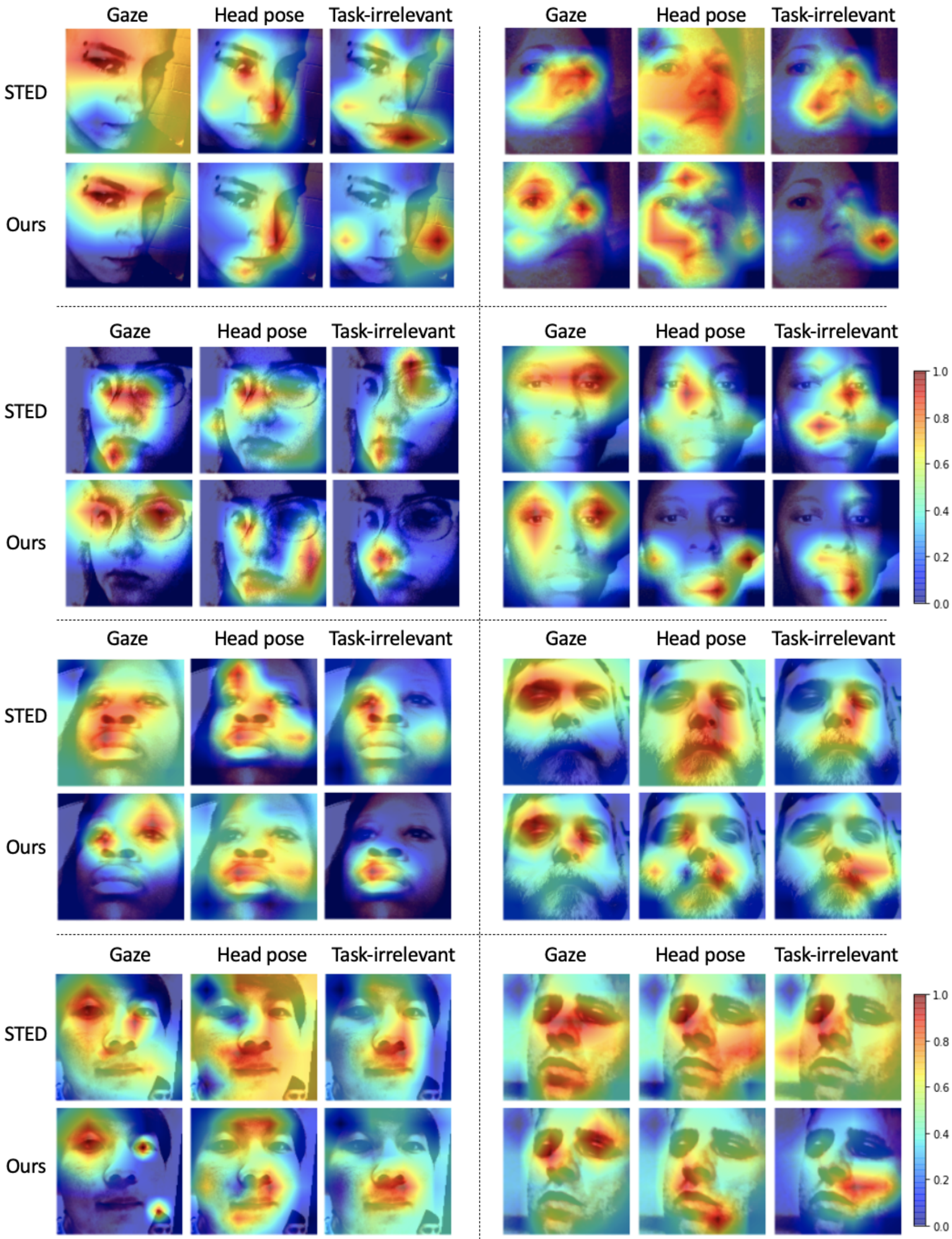

Figure 3: Additional qualitative results on GazeCapture dataset

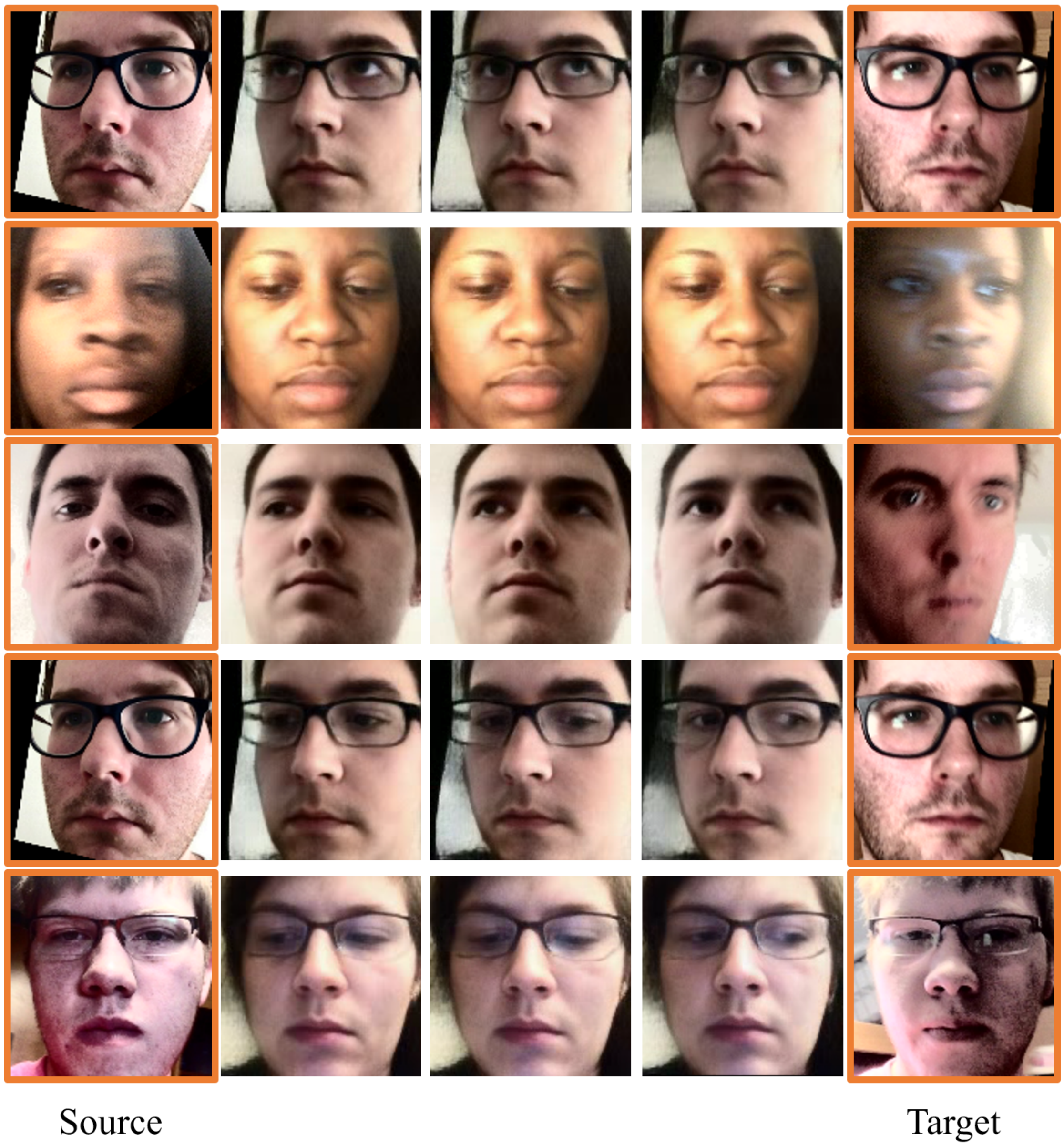

Figure 4: Generated image using the interpolated gaze feature