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 \( R \) in Eq. 1.

\[
\tilde{z}_i^g = R(z_i^g) = R_i^g (R_i^g)^{-1} z_i^g,
\]

where the rotation matrix \( R_i^g \) for the transformation of the source gaze feature \( \tilde{z}_i^g \) is as follows:

\[
R_i^g = \begin{pmatrix}
\cos \phi^g & 0 & \sin \phi^g \\
0 & 1 & 0 \\
-\sin \phi^g & 0 & \cos \phi^g
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 \\
0 & \cos \theta^g & -\sin \theta^g \\
0 & \sin \theta^g & \cos \theta^g
\end{pmatrix}
\]

Here, \((\theta^g, \phi^g)\) represents the pitch and yaw angle of the source gaze direction. Similar to Eq. 2, \( R_i^g \) is defined based on \((\theta^g, \phi^g)\). The redirection process based on these rotation matrices is also applied to head pose \( z_i^h \). Redirected features are used to generate redirected images for supervised learning.

2. Other loss function

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

\[
L_{other} = \|\tilde{x}_i - x_i\|_1 + \sum_{k=1}^{5} \|F_k(\tilde{x}_i) - F_k(x_i)\|_2
\]

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

3. Further Analysis of SG loss

Relationship with contrastive loss. Let’s analyze the operation of \( L_{sg} \) in terms of the well-known contrastive loss. First, \( J_{i,j} \) of \( L_{sg} \) is rearranged using logarithmic and exponential operators as follows:

\[
J_{i,j} = D_{i,j} + \sum_{(i,k)\in N} \log(e^{\delta-D_{i,k}}) + \sum_{(j,l)\in N} \log(e^{\delta-D_{j,l}})
\]

\[
\leq D_{i,j} + \log \sum_{(i,k)\in N} e^{\delta-D_{i,k}} + \log \sum_{(j,l)\in N} e^{\delta-D_{j,l}}
\]

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

\[
L_{sg} = \frac{1}{2|\mathcal{P}|} \sum_{(i,j)\in \mathcal{P}} D_{i,j} + \frac{1}{2|\mathcal{P}|} \sum_i \log \sum_{k} e^{\delta-D_{i,k}}
\]

+ \frac{1}{2|\mathcal{P}|} \sum_j \log \sum_{l} e^{\delta-D_{j,l}}

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 \((z_i^g, z_j^g)\) or \((z_i^h, z_j^h)\) 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 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 \((z_i^g, z_j^g)\), \((z_i^h, z_j^h)\), and \((z_i^g, z_j^h)\) matches the uniform distribution. Therefore, \( z_i^g \) 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

<table>
<thead>
<tr>
<th>Mini-batch</th>
<th>err ( g )</th>
<th>err ( h )</th>
<th>( h \rightarrow g )</th>
<th>( g \rightarrow h )</th>
<th>LPIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>1.973</td>
<td>0.720</td>
<td>1.933</td>
<td>0.334</td>
<td>0.196</td>
</tr>
<tr>
<td>48</td>
<td>2.010</td>
<td>0.770</td>
<td>1.895</td>
<td>0.380</td>
<td>0.203</td>
</tr>
<tr>
<td>64</td>
<td>2.110</td>
<td>0.724</td>
<td>1.993</td>
<td>0.340</td>
<td>0.206</td>
</tr>
<tr>
<td>128</td>
<td>1.964</td>
<td>0.693</td>
<td>1.882</td>
<td>0.330</td>
<td>0.203</td>
</tr>
</tbody>
</table>

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

\[
\text{ContraCAM}_{hw} = \text{Normalize} \left( \text{ReLU} \left( \sum_c \alpha_c A_{hw}^c \right) \right)
\]

\[
\alpha_c = \text{ReLU} \left( \frac{1}{HW} \sum_{h,w} \partial \text{MLP}(z^c/h/w) \right)
\]

where \(A_{hw}^c\) is the feature map or spatial activation extracted from the middle stage (the 6th layer) of encoder \(E\). Also, \(h, w\) and \(c\) indicate the index of height \((H)\), width \((W)\) and channel size \((C)\), respectively. \(\text{Normalize}(x) = \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}\) is a normalization function that maps the range of \(x\) to \([0, 1]\). \(\text{MLP}(\cdot)\) 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 \(\text{MLP}(\cdot)\) (marked as red color) that outputs (continuous) confidence scores. We can see the Pytorch-like pseudocode 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.

Listing 1: Pytorch-style pseudo-code for ContraCAM

```python
1 A = encoder[:2](x) # 6th feature map of encoder
2 z = encoder(x) # feature vector
3 predicted_label = MLP(z)
4 grad = autograd.grad(predicted_label.sum(), A)[0]
5 weight = adaptive_avg_pool2d(grad, output_size=(1, 1))
6 weight = relu(weight) # non-negative clipping
7 # weighted sum
8 cam = sum(weight * A, dim=1, keepdim=True).detach()
9 cam = resize(cam(h, w))
10 cam = normalize(relu(cam)) # normalize to [0, 1]
```

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:

\[
L_{\text{margin}} = \left[ x + y_{i,j} (D_{i,j} - \beta) \right]_+.
\]

Table 2: Performance according to metric loss

<table>
<thead>
<tr>
<th>Metric loss</th>
<th>err ( g )</th>
<th>err ( h )</th>
<th>( h \rightarrow g )</th>
<th>( g \rightarrow h )</th>
<th>LPIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Margin [11]</td>
<td>2.264</td>
<td>0.827</td>
<td>1.994</td>
<td>0.368</td>
<td>0.212</td>
</tr>
<tr>
<td>DSML (tri) [13]</td>
<td>2.100</td>
<td>0.799</td>
<td>1.915</td>
<td>0.377</td>
<td>0.206</td>
</tr>
<tr>
<td>SG (Ours)</td>
<td>1.973</td>
<td>0.720</td>
<td>1.933</td>
<td>0.334</td>
<td>0.196</td>
</tr>
</tbody>
</table>
where the flexible boundary parameter $\beta$ is learnable, and the static margin $\alpha$ 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 $d_S(z_i, z_j) = \frac{\text{var}(x_i - z_i)}{\text{var}(x_i)}$ rather than the Euclidean distance. Here, $\text{var}(z)$ is the variance of $z$. The triplet-based SNR metric loss we adopted is as follows:

$$
\mathcal{L}_{DSML(tr)} = \left[ d_S(z_i^g, z_j^g) - d_S(z_i^s, z_j^h) + \alpha \right]_+ + \left[ d_S(z_i^g, z_j^s) - d_S(z_i^g, z_j^h) + \alpha \right]_+. 
$$

where margin $\alpha$ 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 $z_i^g$ was used instead of $z_i^h$ 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-dataset evaluation protocol. Table 3 shows the performance of the proposed method and other methods on the MPIIGaze, 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 $z_i^g$ to generate a redirected image $\tilde{x}$: $\tilde{x} = G(\text{Concat}(z_i^r, z_i^b, z_i^h))$. 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 $\psi$. 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.

### References


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.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\text{err}_g$</th>
<th>$u \rightarrow g$</th>
<th>$h \rightarrow g$</th>
<th>$\text{err}_h$</th>
<th>$u \rightarrow h$</th>
<th>$g \rightarrow h$</th>
<th>LPIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GazeFlow [12]</td>
<td>5.887</td>
<td>3.778</td>
<td>5.312</td>
<td>3.713</td>
<td>1.714</td>
<td>3.121</td>
<td>0.243</td>
</tr>
<tr>
<td>FAZE [6]</td>
<td>7.312</td>
<td>-</td>
<td>6.714</td>
<td>2.512</td>
<td>-</td>
<td>1.917</td>
<td>0.237</td>
</tr>
<tr>
<td>STED [15]</td>
<td>2.133</td>
<td>0.605</td>
<td>2.312</td>
<td>0.724</td>
<td>0.314</td>
<td>0.442</td>
<td>0.204</td>
</tr>
<tr>
<td>Ours</td>
<td>1.814</td>
<td>0.512</td>
<td>1.994</td>
<td>0.684</td>
<td>0.211</td>
<td>0.339</td>
<td>0.202</td>
</tr>
</tbody>
</table>

(a) MPIIGaze [14]

<table>
<thead>
<tr>
<th>Method</th>
<th>$\text{err}_g$</th>
<th>$u \rightarrow g$</th>
<th>$h \rightarrow g$</th>
<th>$\text{err}_h$</th>
<th>$u \rightarrow h$</th>
<th>$g \rightarrow h$</th>
<th>LPIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GazeFlow† [12]</td>
<td>7.312</td>
<td>-</td>
<td>5.076</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.274</td>
</tr>
<tr>
<td>STED [15]</td>
<td>3.134</td>
<td>0.902</td>
<td>3.307</td>
<td>0.886</td>
<td>0.334</td>
<td>1.002</td>
<td>0.233</td>
</tr>
<tr>
<td>Ours</td>
<td>2.872</td>
<td>0.782</td>
<td>2.902</td>
<td>0.902</td>
<td>0.314</td>
<td>0.987</td>
<td>0.212</td>
</tr>
</tbody>
</table>

(b) Columbia [9]

<table>
<thead>
<tr>
<th>Method</th>
<th>$\text{err}_g$</th>
<th>$u \rightarrow g$</th>
<th>$h \rightarrow g$</th>
<th>$\text{err}_h$</th>
<th>$u \rightarrow h$</th>
<th>$g \rightarrow h$</th>
<th>LPIPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GazeFlow† [12]</td>
<td>17.122</td>
<td>-</td>
<td>3.124</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.264</td>
</tr>
<tr>
<td>STED [15]</td>
<td>13.094</td>
<td>6.413</td>
<td>12.796</td>
<td>0.817</td>
<td>0.662</td>
<td>1.674</td>
<td>0.224</td>
</tr>
<tr>
<td>Ours</td>
<td>11.094</td>
<td>5.498</td>
<td>9.438</td>
<td>0.802</td>
<td>0.403</td>
<td>0.904</td>
<td>0.232</td>
</tr>
</tbody>
</table>

(c) EYEDIAP [1]

Figure 3: Additional qualitative results on GazeCapture dataset
Figure 4: Generated image using the interpolated gaze feature