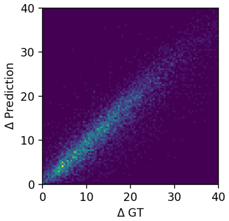
Rebuttal for #12213 | **Self-learning Transformations for Improving Gaze and Head Redirection**

We thank the reviewers for noting that our approach is novel and justified (R1), simple and clearly described (R2), interesting (R1, R3), and that our qualitative results are impressive (R2, R3). We will clarify any open points below.

1) Use of F_d for evaluation (R3). To analyze the effect of using identical F_d models in training and in evaluation, we train a separate ResNet-50 on GazeCapture (training set) to use only for evaluations (results in the tables below). Even with the separate F_d our approach outperforms state-of-the-art methods. This is in line with the trend shown in Tab. 2 of the main paper. We will update the table accordingly in the camera-ready.

	GazeCapture				MPIIFaceGaze				Columbia				EYEDIAP			
	Gaze Redir.	Head Redir.	$g \rightarrow h$	$h \rightarrow g$	Gaze Redir.	Head Redir.	$g \rightarrow h$	$h \rightarrow g$	Gaze Redir.	Head Redir.	$g \rightarrow h$	$h \rightarrow g$	Gaze Redir.	Head Redir.	$g \rightarrow h$	$h \rightarrow g$
StarGAN	4.602	3.989	0.755	3.067	4.488	3.031	0.786	2.783	6.522	3.444	1.029	3.359	14.906	3.929	0.915	4.025
He *et al.*	4.617	1.392	0.560	3.925	5.092	1.372	0.684	3.411	7.345	1.677	0.692	3.831	13.548	1.581	0.663	4.367
Ours	**2.195**	**0.816**	**0.388**	**2.072**	**2.233**	**0.884**	**0.365**	**1.849**	**3.333**	**1.095**	**0.452**	**2.136**	**11.290**	**0.919**	**0.402**	**2.670**

2) How well F_d works (R1). The histogram to the right plots the differences between F_d gaze angle predictions, for randomly sampled image pairs from the GazeCapture test set, against the corresponding ground-truth deltas. The plot indicates a strong correlation between the output from F_d and the ground-truth differences (Pearson corr. coeff. of 0.92). This provides evidence that F_d is a useful choice as an evaluation network and that it can reliably assess changes in gaze direction. We will add this analysis to the camera-ready.



3) Training on other datasets (R1). We train our approach on MPIIFaceGaze and tabulate the cross-dataset results below (similar to Tab. 2 of the main paper). As expected, the performance decreases with MPIIGaze which has much fewer (15) subjects v.s. GazeCapture (993 subjects after pre-processing). However, our performance remains consistently better than state-of-the-art methods across all test datasets.

	GazeCapture					Columbia					EYEDIAP				
	Gaze Redir.	Head Redir.	$g \rightarrow h$	$h \rightarrow g$	LPIPS	Gaze Redir.	Head Redir.	$g \rightarrow h$	$h \rightarrow g$	LPIPS	Gaze Redir.	Head Redir.	$g \rightarrow h$	$h \rightarrow g$	LPIPS
StarGAN	5.684	7.093	0.778	2.906	0.324	7.503	8.031	0.936	2.596	0.469	15.194	7.591	0.772	2.741	0.468
He *et al.*	5.788	2.874	0.755	5.064	0.299	8.156	4.031	0.946	5.197	0.469	16.904	**3.283**	0.696	6.005	0.407
Ours	**3.064**	**2.764**	**0.391**	**1.821**	**0.261**	**3.955**	**3.833**	**0.405**	**1.625**	**0.424**	**14.624**	3.423	**0.308**	**1.648**	**0.362**

4) Balancing of loss terms (R2). While we did not explore all potential combinations of loss term weights, we found that our method is generally robust to how the weights are specifically set. This is evident from the coarse values chosen (and written in line 192 of our submission). In general, balancing the weighted contributions of the sub-objectives given the magnitude of their raw values is a good guideline. Additionally, an emphasis on the reconstruction loss term improves training stability, and a higher functional loss term weight leads to better redirection accuracy. We will provide some experimental results regarding these observations in our final version.

5) Limitations & failure cases (R2, R3). We notice at times that there is difficulty in handling eyeglasses, and that person-specific appearance characteristics (e.g. face shape) are not always preserved fully. In addition, finer details of the face such as moles and freckles are not retained in the output. See 02:45 of our supplementary video submission for an example with eyeglasses. We will expand our discussion of limitations in the final version.

6) Comparison with FAZE (R2). Tab. 1 of our supplementary text shows results of the base model without pseudo-label prediction at the output of the encoder, i.e., the ground-truth is used to rotate the predicted embeddings (as is done in transforming encoder-decoder architectures). This is equivalent to FAZE + a discriminator – we use the same backbone network as FAZE. The addition of the discriminator alone does not improve FAZE. Note that FAZE itself was not proposed as a gaze redirection method and performs poorly, producing images of very low quality.

7) Explanation of factors and conditions in ST-ED (R3). Each factor is composed of an embedding (z) and condition (c) where the condition describes the amount of variation of the embedding (via rotations). This is defined by the transforming encoder-decoder framework where the encoder predicts a rotated embedding, and the decoder takes as input the same embedding, but with a different rotation. As further discussed in the supplementary text (Sec. 3), the conditions can be 1 or 2-dimensional. Head orientations are 2-dimensional in our setting (line 198 of the main paper).

8) Miscellaneous (R1, R3). We will extend the discussion on the relation of our work to deep fakes, and tone down our claim regarding the generalization possibility of ST-ED to other CV tasks (R1). We will improve Fig. 1 as per suggestions, explain the concept of rotational equivariance more clearly, and cite the CONFIG paper (R3).