ReDirTrans: Latent-to-Latent Translation for Gaze and Head Redirection Supplemental Material

Shiwei Jin [1], Zhen Wang [2], Lei Wang [2], Ning Bi [2], Truong Nguyen [1]
[1] ECE Dept. UC San Diego, [2] Qualcomm Technologies, Inc.
{sjin, tqn001}@eng.ucsd.edu, {zhewang, wlei, nbi}@qti.qualcomm.com

A. Overview

In this supplementary document, first, we demonstrate training details when we incorporate ReDirTrans with encoder-generator pair. Then we present some additional quantitative results of ReDirTrans-GAN. Lastly, we introduce the details when we implemented other state-of-the-art baselines.

B. Implementation Details

We trained and evaluated ReDirTrans given two different encoder-generator pairs: 1) the trainable encoder-decoder pair from ST-ED [15]; 2) the pre-trained e4e [13] and StyleGAN [7]. Thus we introduced implementation details in two cases: 1) *ReDirTrans* and 2) *ReDirTrans-GAN*.

B.1. Datasets

GazeCapture [9] is the largest public gaze-related full-face dataset including $1,474$ participants with over two million frames taken under unconstrained scenarios. We utilize its training subset to train the redirector and evaluate the redirection precision with its test subset.

MPIIFaceGaze [14] is a widely used benchmark dataset for the in-the-wild gaze estimation task. It includes $37,667$ full-face images captured from 15 participants with varied head orientations, multiple gaze directions and different illuminations. We utilize this dataset to evaluate the cross-dataset redirection performance.

CelebA-HQ [6] is a high-quality version of CelebA [10] that consists of $30,000$ images at 1024×1024 resolution. We utilize this dataset for evaluating the cross-dataset qualitative redirection performance.

B.2. Preprocessing Steps

1) *ReDirTrans*: We preprocessed the image data to acquire a 128×128 restricted range of face images aligned with key points of the nose and eyes, followed by [15].

2) *ReDirTrans-GAN*: We preprocessed the image data to acquire 256×256 full-face images aligned with key points of the mouth and eyes. We utilized reflective padding to the blank areas after alignment and then covered these areas with Gaussian blur, followed by [6].

ReDirTrans		Layers/Blocks
P	Pseudo Label Branch	FC(3072, 96, w/bias), LeakyReLU() FC(96, 4, w/bias), pi/2*Tanh()
	Embedding Branch	FC(3072, 3072, w/bias), LeakyReLU() FC(3072, 96, w/bias)
	DP	FC(96, 1024, w/bias), LeakyReLU() FC(1024, 3072, w/bias)

Table 1. The architecture of the projector and deprojector in **ReDirTrans**. **P** denotes projector and **DP** denotes deprojector.

B.3. Projector-Deprojector

Since the inputs to the projector have already been the decoded latent vectors from images, we utilized several fully connected modules as the architectures of the projector-deprojector. 1) *ReDirTrans*: Unlike ST-ED projecting the input latent vector into nine attribute embeddings, our proposed ReDirTrans only projected it into the aimed attribute (gaze directions and head orientations) embeddings. The size of the estimated label and embedding of one attribute are 2 and 3×16, respectively. The details are illustrated in Table 1.

1) *ReDirTrans-GAN*: As for the ReDirTrans-GAN, the main difference comes from the size of latent vectors in latent space $\mathcal{F}$. The details are shown in Table 2.

B.4. Encoder-Generator and Loss Functions

1) *ReDirTrans*: ST-ED proposed the architecture of encoder-decoder pair given the DenseNet [4] architecture, for 128×128 output. As for the decoder, the convolutional layers were replaced by the transposed convolutional layers and the average-pooling layers. The detailed encoder-decoder structure is illustrated in [15]. To ensure the generation quality, we utilized a PatchGAN [5] discriminator with corresponding adversarial loss as proposed in ST-ED during the training.

ReDirTrans-GAN		Layers/Blocks
P	Pseudo Label Branch	FC(512, 64, w/bias), LeakyReLU() FC(64, 4, w/bias), pi/2*Tanh()
	Embedding Branch	FC(512, 128, w/bias), LeakyReLU() FC(128, 96, w/bias)
	DP	FC(96, 256, w/bias), LeakyReLU() FC(256, 512, w/bias)

Table 2. The architecture of the projector and deprojector in **ReDirTrans-GAN**. **P** denotes projector and **DP** denotes deprojector.

2) *ReDirTrans-GAN*: Since both e4e and StyleGAN were pretrained and fixed during the training, the image discriminator mentioned above was no longer used. Instead, to maintain the perceptual quality and editability of latent codes after redirection as the original latent codes encoded by e4e, we kept utilizing the e4e proposed delta-regularization loss $\mathcal{L}_{d-reg}$ and the adversarial loss $\mathcal{L}_{adv}$ by a latent discriminator [11]. Noted that we applied these two loss functions to the modified latent vectors after redirection to maintain the editability of e4e encoded latent vectors.

B.5. Gaze and Head Pose Estimation Network

During training, we need a pretrained gaze and head pose estimation network $\xi_{hg}(\cdot)$ as the estimator to supervise the redirection process. During the evaluation, we require another different external pretrained gaze and head pose estimation network $\xi'_{hg}(\cdot)$, which is unseen during training, to evaluate the consistency of the aimed attributes between redirected and target samples. We followed the pipeline proposed by ST-ED, which utilized a VGG-16-based $\xi_{hg}(\cdot)$ [12] and a ResNet50-based $\xi'_{hg}(\cdot)$ [3].

1) *ReDirTrans*: We retrained and employed the VGG-16-based $\xi_{hg}(\cdot)$ and ResNet50-based $\xi'_{hg}(\cdot)$ as the gaze and head pose estimators, given the architectures and training parameters illustrated in [15].

2) *ReDirTrans-GAN*: To fit the different sizes of input (256×256) and output images (1024×1024) with the full face range when training and evaluating ReDirTrans-GAN, we downsampled the output images to 256×256. The fully connected modules after the convolutional part of $\xi_{hg}(\cdot)$ and $\xi'_{hg}(\cdot)$ were modified accordingly for different input sizes compared with the ST-ED version. The detailed structures of $\xi_{hg}(\cdot)$ and $\xi'_{hg}(\cdot)$ are shown in Table 3 and Table 4, respectively.

B.6. Training Hyperparameters

1) *ReDirTrans*: We trained ReDirTrans and the encoder-decoder pair with the same hyperparameters as ST-ED [15] by using over 1.4×10^6 full-face images from GazeCapture Training subset.

Nr.	layers / blocks
0	VGG-16 Conv layers
1	AvgPool2d(size=4, stride=4)
2	FC(2048, 128, w/bias), LeakyReLU()
3	FC(128, 64, w/bias), LeakyReLU()
4	FC(64, 4, w/bias), $0.5\pi \cdot \tanh()$

Table 3. The architecture of the VGG-16-based gaze direction and head orientation estimation network, $\xi_{hg}(\cdot)$.

Nr.	layers / blocks
0	ResNet-50 Conv layers, stride of MaxPool2d=1
1	FC(2048, 4, w/bias)

Table 4. The architecture of the ResNet50-based gaze direction and head orientation estimation network, $\xi'_{hg}(\cdot)$.

2) *ReDirTrans-GAN*: We randomly chose $10,000$ images from the GazeCapture training subset to train ReDirTrans-GAN since both the encoder and generator are fixed and pretrained. The number of epochs is 2 with a batch size of 2. The initial learning rate is 10^{-4} and is decayed by 0.8 every $3,000$ iterations. The optimizer is Adam [8] with the default momentum value of $\beta_1 = 0.9, \beta_2 = 0.999$.

The loss weights are $\lambda_r = 8$, $\lambda_L = 8$, $\lambda_{ID} = 5$, $\lambda_a = 1$, $\lambda_l = 5$, $\lambda_e = 2$, $\lambda_p = 10$, $\lambda_{d-reg} = 0.0002$, $\lambda_{adv} = 2$, where λ_{d-reg} and λ_{adv} are the weights of delta-regularization loss $\mathcal{L}_{d-reg}$ and the adversarial loss $\mathcal{L}_{adv}$, respectively.

B.7. Redirection Step

We applied rotation matrices built by the pitch and yaw to the estimated embeddings for redirection purposes.

$$\begin{bmatrix} \cos\phi^i & 0 & \sin\phi^i \\ 0 & 1 & 0 \\ -sin\phi^i & 0 & \cos\phi^i \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta^i & -\sin\theta^i \\ 0 & \sin\theta^i & cos\theta^i \end{bmatrix}, i \in \{1, 2\} \tag{1}$$

where ϕ represents yaw and θ represents pitch, and index $i \in \{1, 2\}$ represents gaze directions and head orientations, respectively.

C. Further Results

C.1. Gaze Correction

Table 5 and Table 6 present within- and cross-dataset evaluation performance for gaze correction tasks. e4e inversion results can maintain gaze directions and head orientations better in CelebA-HQ than GazeCapture since samples in CelebA-HQ have much less varied gaze directions and head orientations. However, after we included ReDirTrans in the

	Gaze Redir ↓	Head Redir ↓	ID (I_t) ↓	ID ($\hat{I}_t$) ↓
e4e	11.302	4.13	0.377	–
ReDirTrans-GAN	**2.505**	**1.020**	0.388	0.128

Table 5. Within-dataset gaze correction performance given the input latent vectors encoded by e4e in the GazeCapture test subset. As for the ID similarity, we compared the redirected image with the real target image (I_t) and its inverted image ($\hat{I}_t$).

	Gaze Redir ↓	Head Redir ↓	ID (I_t) ↓	ID ($\hat{I}_t$) ↓
e4e	4.448	2.586	0.286	–
ReDirTrans-GAN	**3.157**	**2.257**	0.314	0.099

Table 6. Corss-dataset gaze correction performance given the input latent vectors encoded by e4e in CelebA-HQ. As for the ID similarity, we compared the redirected image with the real target image (I_t) and its inverted image ($\hat{I}_t$).

	Gaze Redir ↓	Head Redir ↓	ID (I_t) ↓	ID ($\hat{I}_t$) ↓
ReDirTrans-GAN	2.648	1.863	0.212	0.130

Table 7. Within-dataset gaze and head redirection performance given the input latent vectors encoded by e4e in the GazeCapture test subset. As for the ID similarity, we compared the redirected image with the real target image (I_t) and its inverted image ($\hat{I}_t$).

inversion pipeline as ReDirTrans-GAN, we can successfully maintain gaze directions and head orientations without affecting identity information (ID), which was measured by a pretrained ArcFace model [2]. Fig. 1 shows several examples.

C.2. Redirection Accuracy of ReDirTrans-GAN

Table 7 presents the redirection accuracy of ReDirTrans-GAN in the GazeCapture test subset. We can observe that ReDirTrans-GAN cannot achieve as accurate redirection performance as ReDirTrans, which worked with the trainable encoder-decoder pair. There exists a **trade-off** between redirection accuracy and the following considerations:

- We utilized fixed encoder and generator parameters during the redirector training to ensure no modification to the predefined latent space;
- e4e encoded latent vectors in W^+ have limitations to understanding gaze in Section 4.7. e4e was trained with the FFHQ dataset, which does not include samples with as varied gaze directions and head orientations as the samples in GazeCapture. Given some cases with large gaze directions or head orientations, e4e cannot invert them very well;
- We kept the redirected latent codes within the 'high editability space' proposed by e4e to allow for further editing with other face editing techniques, sacrificing some quality (redirection accuracy);
- Extended face covering ranges and down-sampling of high-resolution generated images could cause the performance drop.
- The deprojector learned that the redirected latent vectors after addition and subtraction operations would not deviate away from the original input latent vectors. Thus ReDirTrans-GAN cannot redirect some extreme cases as well as ReDirTrans did, especially for head orientations.
- Predefined face alignments (four eyes corners and two mouth corners) restricts both the encoder and generator's ability for extreme head pose synthesis.

In summary, we made a deliberate choice to use a fixed encoder-generator pair, preserve edited latent codes in W^+, and edit within the 'high editability space' to maintain compatibility with continuing facial attribute editing by other methods. ReDirTrans-GAN provides a solution to edit attributes in predefined feature spaces that have limited abilities to depict those attributes. It also addresses the face editing task of redirecting or correcting gaze from the latent code perspective. Fig. 2 and Fig. 3 show redirected samples with modifying gaze directions and head orientations separately.

C.3. Layer-wise Weights

Given the layer-wise representation of latent vectors in W^+ space, we proposed layer-wise weights loss to measure the contribution of each layer for the corresponding attribute redirection. We compared the redirected samples with and without considering the layer-wise weights loss, shown in Fig. 4. We observed that gaze directions and head orientations become entangled without this loss and the network tends to learn a specific combination of gaze directions

Nr.	layers / blocks
0	FC(4, 32, w/bias), LeakyReLU()
1	FC(32, 64, w/bias), LeakyReLU()
2	FC(64, 64, w/bias), LeakyReLU()
3	FC(64, 4, w/bias)

Table 8. Architecture of the scale transformation network. This network transforms the pseudo labels into the scalars for the multiplication with the the global latent directions in $\mathcal{F}$.

and head orientations. However, when we utilized this loss with the estimated layer-wise weights to modify each layer's output residuals further. In that case, gaze directions and head orientations can be disentangled and redirected independently.

D. State-of-the-Art Baselines

D.1. ST-ED

We preprocessed the face images as shown in ST-ED based on code repository [1]. We trained and evaluated ST-ED based on their published code repository [2].

D.2. VecGAN

VecGAN [1] was originally proposed to modify facial attributes, such as gender, age, hair color, smiling and bangs. We applied the VecGAN's addition and subtraction operations in the latent space $\mathcal{F}$ to the ST-ED proposed encoder-decoder pair. Instead of estimating embeddings from input latent vectors, VecGAN only estimates the pseudo labels of the aimed attributes. We utilized four orthogonal global latent directions 1×1344 for pitch and yaw control of the gaze directions and head orientations separately. We cannot directly apply the estimated labels to the global latent directions in $\mathcal{F}$ by the multiplication operation due to the scale problem. Thus we applied a shallow network for transforming pseudo labels to scalars to fit for the global latent directions. The network structure is shown in Table 8.

[1] https://github.com/swook/faze_preprocess
[2] https://github.com/zhengyuf/STED-gaze

Figure 1. Gaze Correction Samples in CelebA-HQ. We set the same image as input and target to redirect the wrong gaze directions and head orientations given e4e inversion results.

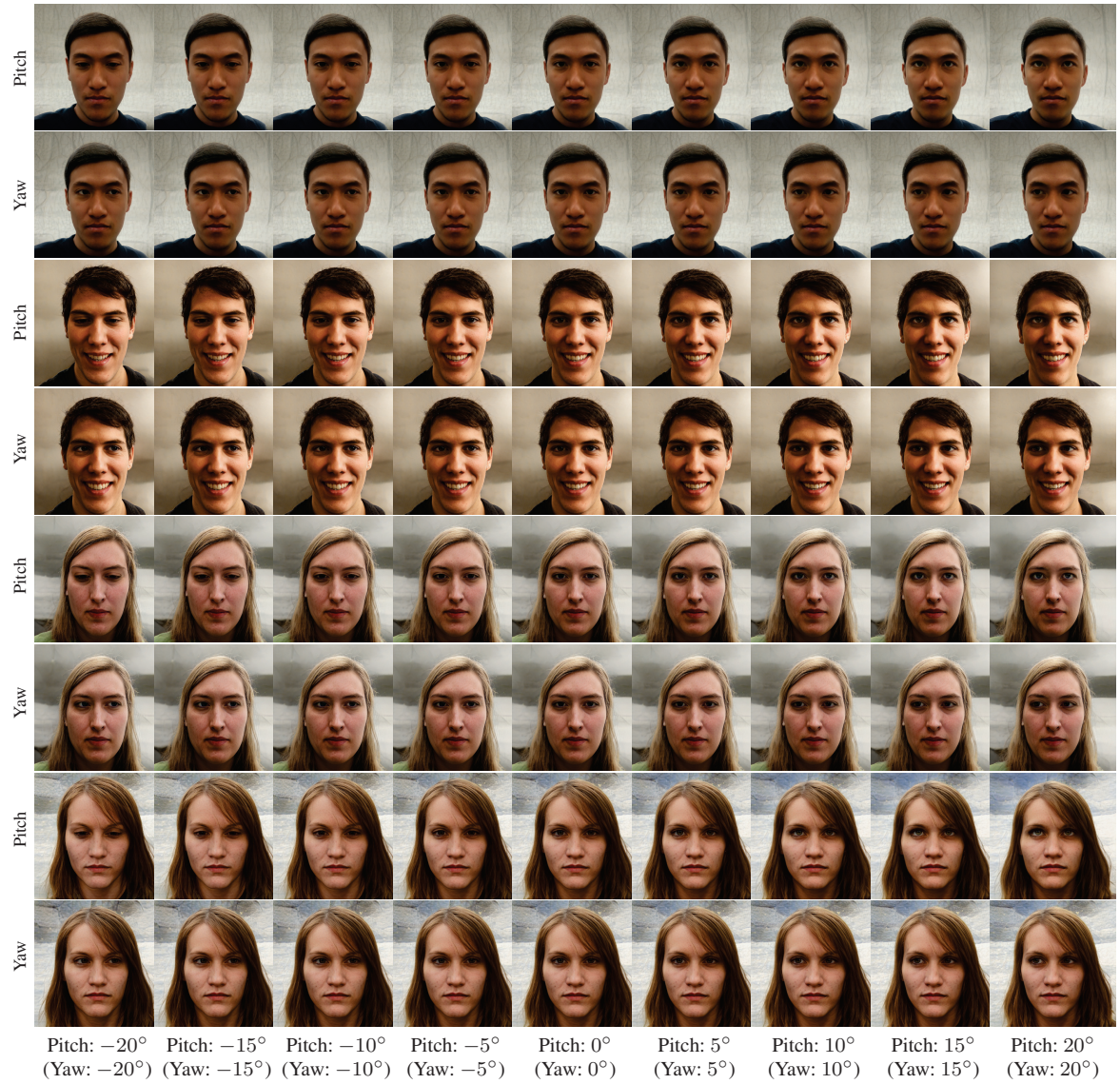

Figure 2. Gaze Redirection Results. The head orientations are all set as $(0^\circ, 0^\circ)$. 'Pitch' means that we only redirect the pitch component of gaze directions and set yaw as 0°. 'Yaw' means that we only redirect the yaw component of gaze directions and set pitch as 0°. The redirected angles are listed at the bottom of the figure.

Figure 3. Head Redirection Results. The gaze directions are all set as $(0^\circ, 0^\circ)$. 'Pitch' means that we only redirect the pitch component of head orientations and set yaw as 0°. 'Yaw' means that we only redirect the yaw component of head orientations and set pitch as 0°. The redirected angles are listed at the bottom of the figure.

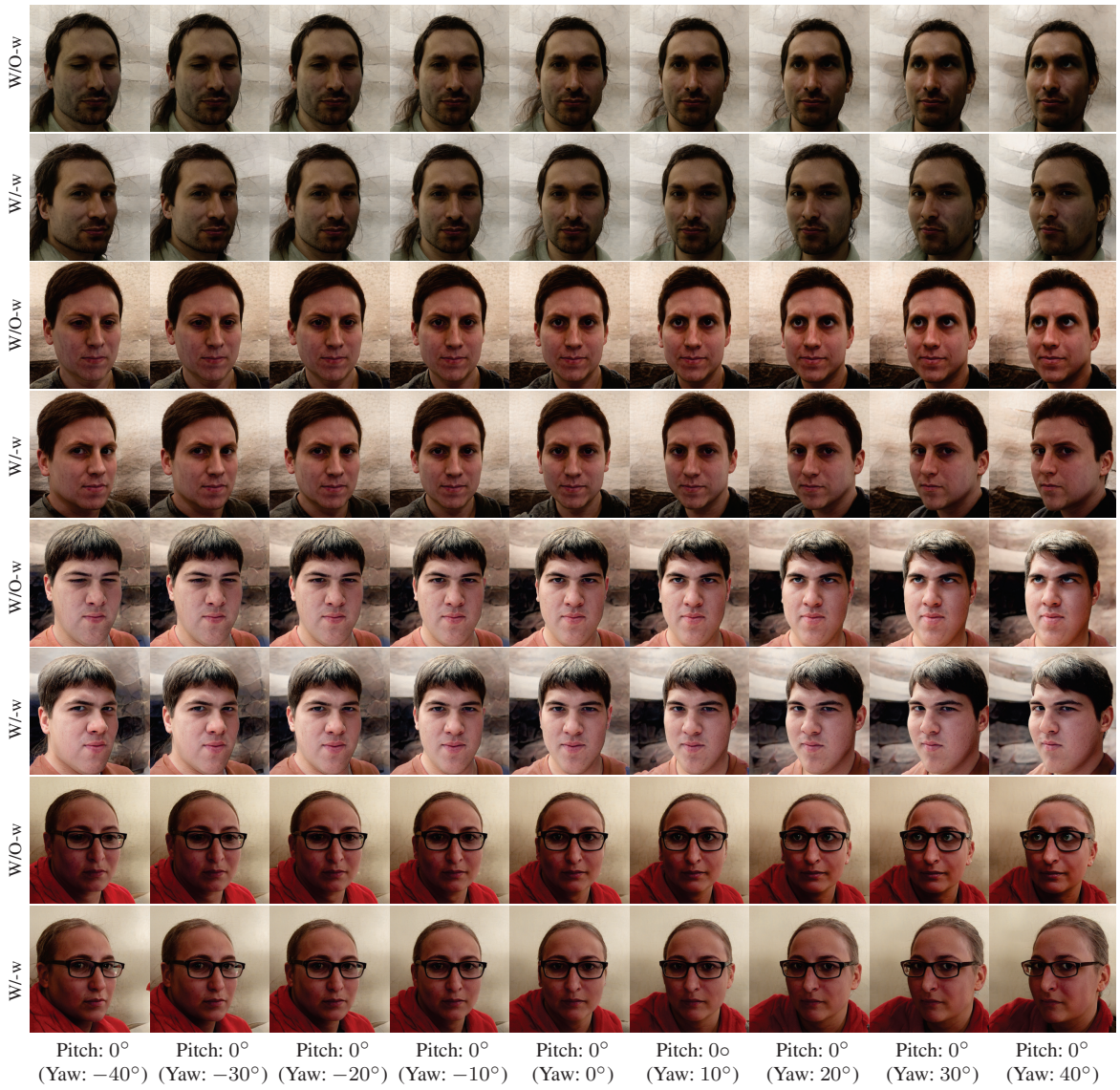

Figure 4. The comparison of redirected samples with or without using layer-wise weights loss in ReDirTrans-GAN. The gaze directions are all set as $(0^\circ, 0^\circ)$. We only redirect the head orientations given the provided pitch and yaw values below the figure. 'W/-w' denotes the redirected samples with the layer-wise weights loss. 'W/O-w' denotes the redirected samples without the layer-wise weights loss.

References

[1] Yusuf Dalva, Said Fahri Altindis, and Aysegul Dundar. Vecgan: Image-to-image translation with interpretable latent directions. *arXiv preprint arXiv:2207.03411*, 2022. 4

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[9] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016. 1

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 1

[11] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *arXiv preprint arXiv:2005.07728*, 2020. 2

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[13] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 1

[14] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–60, 2017. 1

[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*, 33:13127–13138, 2020. 1, 2