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

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

Abstract

Learning-based gaze estimation methods require large amounts of training data with accurate gaze annotations. Facing such demanding requirements of gaze data collection and annotation, several image synthesis methods were proposed, which successfully redirected gaze directions precisely given the assigned conditions. However, these methods focused on changing gaze directions of the images that only include eyes or restricted ranges of faces with low resolution (less than 128×128) to largely reduce interference from other attributes such as hairs, which limits application scenarios. To cope with this limitation, we proposed a portable network, called ReDirTrans, achieving latent-to-latent translation for redirecting gaze directions and head orientations in an interpretable manner. ReDirTrans projects input latent vectors into aimed-attribute embeddings only and redirects these embeddings with assigned pitch and yaw values. Then both the initial and edited embeddings are projected back (deprojected) to the initial latent space as residuals to modify the input latent vectors by subtraction and addition, representing old status removal and new status addition. The projection of aimed attributes only and subtraction-addition operations for status replacement essentially mitigate impacts on other attributes and the distribution of latent vectors. Thus, by combining ReDirTrans with a pretrained fixed e4e-StyleGAN pair, we created ReDirTrans-GAN, which enables accurately redirecting gaze in full-face images with 1024×1024 resolution while preserving other attributes such as identity, expression, and hairstyle. Furthermore, we presented improvements for the downstream learning-based gaze estimation task, using redirected samples as dataset augmentation.

1. Introduction

Gaze is a crucial non-verbal cue that conveys attention and awareness in interactions. Its potential applications include mental health assessment [5,18], social attitudes analysis [19], human-computer interaction [12], automotive assistance [30], AR/VR [6,34]. However, developing a robust unified learning-based gaze estimation model requires large amounts of data from multiple subjects with precise gaze annotations [42,44]. Collecting and annotating such an appropriate dataset is complex and expensive. To overcome this challenge, several methods have been proposed to redirect gaze directions [17,39,41,42,44] in real images with assigned directional values to obtain and augment training data. Some works focused on generating eye images with new gaze directions by either 1) estimating warping maps [41,42] to interpolate pixel values or 2) using encoder-generator pairs to generate redirected eye images [17,39].

ST-ED [44] was the first work to extend high-accuracy gaze redirection from eye images to face images. By disentangling several attributes, including person-specific appearance, it can explicitly control gaze directions and head orientations. However, due to the design of the encoder-decoder structure and limited ability to maintain appearance features by a 1×1024 projected appearance embedding, ST-ED generates low-resolution (128×128) images with restricted face range (no hair area), which narrows the application ranges and scenarios of gaze redirection.

As for latent space manipulation for face editing tasks, large amounts of works [2–4, 14, 31, 35] were proposed to modify latent vectors in predefined latent spaces (W [21], W^+ [1] and S [40]). Latent vectors in these latent spaces can work with StyleGAN [21, 22] to generate high-quality and high-fidelity face images with desired attribute editing. Among these methods, Wu *et.al* [40] proposed the latent space S working with StyleGAN, which achieved only one degree-of-freedom gaze redirection by modifying a certain channel of latent vectors in S by an uninterpreted value instead of pitch and yaw values of gaze directions.

Considering these, we proposed a new method, called ReDirTrans, to achieve latent-to-latent translation for redirecting gaze directions and head orientations in high-resolution full-face images based on assigned directional values. Specifically, we designed a framework to project input latent vectors from a latent space into the aimed-attribute-only embedding space for an interpretable redirection process. This embedding space consists of estimated pseudo conditions and embeddings of aimed attributes, where conditions describe deviations from the canonical status and embeddings are the 'carriers' of the conditions. In this embedding space, all transformations are implemented by rotation matrices multiplication built from pitch and yaw values, which can make the redirection process more interpretable and consistent. After the redirection process, the original embeddings and redirected ones are both decoded back to the initial latent space as the residuals to modify the input latent vectors by subtraction and addition operations. These operations represent removing the old state and adding a new one, respectively. ReDirTrans only focuses on transforming embeddings of aimed attributes and achieves status replacement by the residuals outputted from weight-sharing deprojectors. ReDirTrans does not project or deproject other attributes with information loss; and it does not affect the distribution of input latent vectors. Thus ReDirTrans can also work in a predefined feature space with a fixed pretrained encoder-generator pair for the redirection task in desired-resolution images.

In summary, our contributions are as follows:

- A latent-to-latent framework, *ReDirTrans*, which projects latent vectors to an embedding space for an interpretable redirection process on aimed attributes and maintains other attributes, including appearance, in initial latent space with no information loss caused by projection-deprojection processes.
- A portable framework that can seamlessly integrate into a pretrained GAN inversion pipeline for high-accuracy redirection of gaze directions and head orientations, without the need for any parameter tuning of the encoder-generator pairs.
- A layer-wise architecture with learnable parameters that works with the fixed pretrained StyleGAN and achieves redirection tasks in high-resolution full-face images through *ReDirTrans-GAN*.

2. Related Works

Gaze and Head Redirection. Methods for redirecting gaze directions can be broadly classified into two categories: warping-based methods and generator-based methods. Deepwarp [13, 23] presented a deep network to learn warping maps between pairs of eye images with different gaze directions, which required large amounts of data with annotations. Yu *et al.* [41] utilized a pretrained gaze estimator and synthetic eye images to reduce the reliance on annotated real data. Yu *et al.* [42] further extended the warping-based methods in an unsupervised manner by adding a gaze representation learning network. As for the generator-based methods, He *et al.* [17] developed a GAN-based network for generating eye images with new gaze directions. FAZE [26] proposed an encoder-decoder architecture to transform eye images into latent vectors for redirection with rotation matrix multiplication, and then decode the edited ones back to the synthetic images with new gaze directions. ST-ED [44] further extended the encoder-decoder pipeline from gaze redirection only to both head and gaze redirection over full face images by disentangling latent vectors, and achieving precise redirection performance. However, ST-ED generates images with a restricted face range (no hair area) with a size of 128×128. We further improve the redirection task by covering the full face range with 1024×1024 resolution.

Latent Space Manipulation. Numerous methods investigated the latent space working with StyleGAN [21, 22] to achieve semantic editing in image space due to its meaningful and highly disentangled properties. As for the supervised methods, InterFaceGAN [31] determined hyperplanes for the corresponding facial attribute editing based on provided labels. StyleFlow [2] proposed mapping a sample from a prior distribution to a latent distribution conditioned on the target attributes estimated by pretrained attribute classifiers. Given the unsupervised methods, GANSpace [15], SeFa [32] and TensorGAN [14] leveraged principal components analysis, eigenvector decomposition and higher-order singular value decomposition to discover semantic directions in latent space, respectively. Other self-supervised methods proposed mixing of latent codes from other samples for local editing [8, 9], or incorporating the language model CLIP [29] for text-driven editing [27].

Domain Adaptation for Gaze Estimation. Domain gaps among different datasets restrict the application range of pretrained gaze estimation models. To narrow the gaps, a few domain adaptation approaches [37, 38] were proposed for the generic regression task. SimGAN [33] proposed an unsupervised domain adaptation method for narrowing the gaps between real and synthetic eye images. HGM [36] designed a unified 3D eyeball model for eye image synthesis and cross-dataset gaze estimation. PnP-GA [25] presented a gaze adaptation framework for generalizing gaze estimation in new domains based on collaborative learning. Qin *et al.* [28] utilized 3D face reconstruction to rotate head orientations together with changed eye gaze accordingly to enlarge overlapping gaze distributions among datasets. These adaptation methods typically rely on restricted face or eye images to alleviate interference from untargeted attributes. Our work incorporates the redirection task in a predefined meaningful feature space with controllable attributes to achieve high-resolution and full-face redirection.

3. Method

3.1. Problem Statements

Our goal is to train a conditional latent-to-latent translation module for face editing with physical meaning, and it can work either with a trainable or fixed encoder-generator

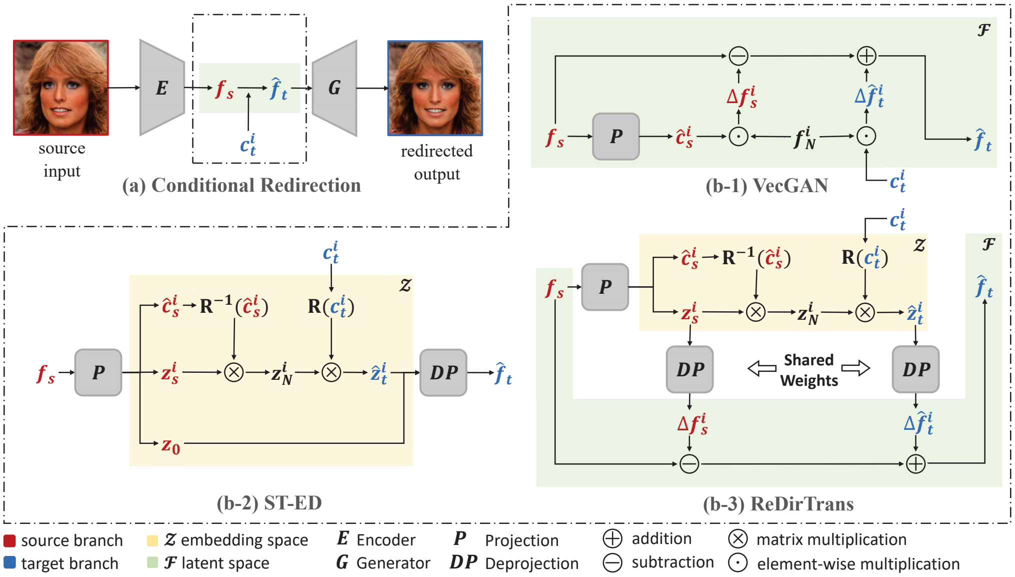

Figure 1. Conditional redirection pipeline and comparison among different redirectors. (a) We first encode the input image into the latent vector. Given the provided conditions, we modify the latent vector and send it to a generator for image synthesis with only aimed attribute redirection. (b) We compared our proposed redirector with two state-of-the-art methods by omitting common modules (basic encoders and decoders/generators) and focusing on the unique components: (b-1) VecGAN achieves editing in feature space $\mathcal{F}$ given projected conditions from latent vectors with a global direction f_N^i. (b-2) ST-ED projects the latent vector into conditions and embeddings of aimed attributes, and one appearance-related high dimensional embedding z_0 in embedding space $\mathcal{Z}$. After interpretable redirection process in space $\mathcal{Z}$, all embeddings are concatenated and projected back to space $\mathcal{F}$. (b-3) Our proposed ReDirTrans projects the latent vector into conditions and embeddings of aimed attributes only. After an interpretable redirection process, both original and redirected embeddings are deprojected back to initial space $\mathcal{F}$ as residuals. These residuals modify the input latent vector by subtraction and addition operations, which represent the initial status removal and the new status addition, respectively. This approach efficiently reduces effects on other attributes (especially the appearance related information) with fewer parameters than ST-ED.

pair. This editing module first transforms input latent vectors from encoded feature space $\mathcal{F}$ to an embedding space $\mathcal{Z}$ for redirection in an interpretable manner. Then it deprojects the original and redirected embeddings back to the initial feature space $\mathcal{F}$ for editing input latent vectors. The edited latent vectors are fed into a generator for image synthesis with the desired status of aimed facial attributes. The previous GAN-based work [10,14,31] achieved a certain facial attribute editing with a global latent residual multiplied by a scalar without physical meaning to describe the relative deviation from the original status. To make the whole process interpretable and achieve redirection directly based on the new gaze directions or head poses, we follow the assumption proposed by [44], where the dimension of an embedding is decided by the corresponding attribute's degree of the freedom (DoF) and redirection process is achieved by the rotation matrices multiplication. Thus the transformation equivariant mappings can be achieved between the embedding space $\mathcal{Z}$ and image space. To be specific, normalized gazes or head poses can be represented by a two-dimensional embedding with the pitch and yaw as the controllable conditions. The embeddings can be edited (multiplied) by the rotation matrices built from the pitch and yaw for achieving redirection (rotation) of aimed attributes in image space accordingly through our proposed redirector.

3.2. Redirector Architecture

ST-ED is one of the state-of-the-art architectures for gaze and head poses redirection over face images [44] shown in Fig. 1 (b-2). ST-ED projects the input latent vector f to non-varying embeddings z^0 and M varying ones with corresponding estimated conditions $\{(z^i, \hat{c}^i)|i \in [1, M]\}$, where $\hat{c}^i$ describes the estimated amount of deviation from the canonical status of the attribute i, and it can be compared with the ground truth c^i for the learning of conditions from latent vectors. The non-varying embedding z^0 defines subject's appearance, whose dimension is much larger (over twenty times larger in ST-ED) than other varying embeddings. It is inefficient to project input latent vectors into a high-dimensional embedding to maintain non-varying information such as identity, hairstyle, etc. Thus, we propose a new redirector architecture, called *ReDirTrans*, shown in Fig. 1 (b-3), which transforms the source latent vector f_s to the embeddings of aimed attributes through the projec-

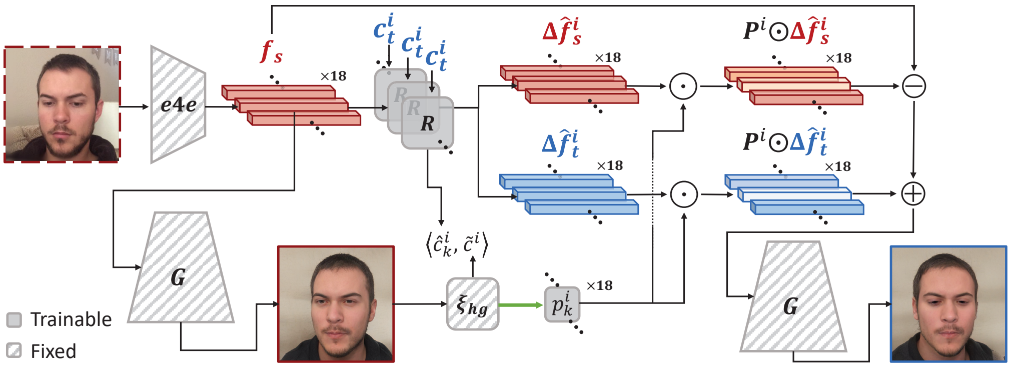
e4e
f_s
×18
c_t^i
R
$\Delta\hat{f}_s^i$
$\Delta\hat{f}_t^i$
$P^i \odot \Delta\hat{f}_s^i$
$P^i \odot \Delta\hat{f}_t^i$
$\langle \hat{c}_k^i, \tilde{c}^i \rangle$
G
ξ_{hg}
p_k^i
▢ Trainable
▨ Fixed

Figure 2. ReDirTrans-GAN: Layer-wise ReDirTrans with fixed e4e and StyleGAN. Given the multi-layer representation of latent vectors in $W^+ \subseteq \mathbb{R}^{18\times512}$, we feed each layer into an independent ReDirTrans for redirection task given the provided target condition c_t^i, where i represents a certain attribute. $\tilde{c}^i$ denotes the pseudo condition estimated directly from the inverted image by a pretrained HeadGazeNet ξ_{hg} [44]. We calculate errors between estimated conditions $\hat{c}_k^i, k \in [1, 18]$ from multiple ReDirTrans and $\tilde{c}^i$ for supervising the trainable weights learning (green arrow) based on the Layer-wise Weights Loss described in Eq. 10 to decide which layers should contribute more to a certain attribute redirection. Given the estimated weights and initial latent vectors f_s, we can acquire the final disentangled latent vector $\hat{f}_{t,d}$ based on Eq. 2 for redirected samples synthesis.

tor $\boldsymbol{P}$ and redirects them given the newly provided target conditions c_t. Then we deproject both original embeddings $\mathbf{z_s}$ and redirected embeddings $\hat{\mathbf{z}}_\mathbf{t}$ back to the feature space $\mathcal{F}$ through the weights-sharing deprojectors $\boldsymbol{DP}$ to acquire latent residuals. These residuals contain source and target status of aimed attributes, denoted as Δf_s^i and $\Delta\hat{f}_t^i$, respectively. Inspired by addition and subtraction [10, 31] for face editing in feature space $\mathcal{F}$, the edited latent vector is

$$\hat{f}_t = f_s + \sum_{i=1}^{M}(-\Delta f_s^i + \Delta\hat{f}_t^i), i \in [1, M], \tag{1}$$

where the subtraction means removing source status and the addition indicates bringing in new status. The projector $\boldsymbol{P}$ ensures that the dimension of embeddings can be customized based on the degrees of freedom of desired attributes, and the transformations can be interpretable with physical meanings. The deprojector $\boldsymbol{DP}$ enables the original and edited features in the same feature space, allowing ReDirTrans to be compatible with pretrained encoder-generator pairs that are typically trained together without intermediate (editing) modules. ReDirTrans reduces parameters by skipping projection (compression) and deprojection (decompression) of the features that are not relevant to the desired attributes, but vital for final image synthesis.

3.3. Predefined Feature Space

Except for the *trainable* encoder-decoder (or -generator) pair to learn a specific feature space for redirection task as ST-ED did, ReDirTrans can also work in the predefined feature space to coordinate with *fixed*, *pretrained* encoder-generator pairs. For our implementation, we chose the $W^+ \in \mathbb{R}^{18\times512}$ feature space [1], which allows us to utilize StyleGAN [22] for generating high-quality, high-fidelity face images. We refer to this implementation as *ReDirTrans-GAN*. Considering multi-layer representation of the latent vector [1] and its semantic disentangled property between different layers [2, 15] in W^+ space, we proposed layer-wise redirectors, shown in Fig. 2, rather than using a single ReDirTrans to process all (18) layers of the latent vector. To largely reduce the interference between different layers during redirection, we assume that if one attribute's condition can be estimated from certain layers with less errors than the others, then we can 'modify' these certain layers with higher weights $p_k^i, k \in [1, 18]$ than others to achieve redirection of the corresponding attribute i only. $\boldsymbol{P}^i = [p_1^i, \cdots, p_{18}^i]^T \in \mathbb{R}^{18\times1}$, as part of network parameters, is trained given the loss function described in Eq. 10. The final disentangled latent vectors after redirection is

$$\hat{f}_{t,d} = f_s + \sum_{i=1}^{M}\boldsymbol{P}^i \odot (-\Delta f_s^i + \Delta\hat{f}_t^i), i \in [1, M], \tag{2}$$

where $\odot$ means element-wise multiplication and $(-\Delta f_s^i + \Delta\hat{f}_t^i) \in \mathbb{R}^{18\times512}$. One **challenge** regarding the predefined feature space comes from the inversion quality. There exist attribute differences between input images and inverted results, shown in Fig. 4 and 6, which means that the conditions in source images cannot be estimated from source latent vectors. To solve this, instead of using conditions from source images, we utilized estimated conditions from the inverted images, which ensures the correctness and consistence of conditions learning from latent vectors.

3.4. Training Pipeline

Given a pair of source and target face images, I_s and I_t from the same person, we utilize an encoder to first transform I_s into the feature space $\mathcal{F}$, denoted as f_s. We further disentangle f_s into the gaze-direction-related embedding z_s^1 and the head-orientation-related embedding z_s^2 with corresponding estimated conditions: $\hat{c}_s^1$ and $\hat{c}_s^2$ by the projector $\boldsymbol{P}$. Then we build rotation matrices using the pitch and yaw from estimated conditions $(\hat{c}_s^1, \hat{c}_s^2)$ and target conditions (c_t^1, c_t^2) to normalize embeddings and redirect them to the new status, respectively:

$$\begin{aligned} &\text{Normalization: } z_N^i = \mathbf{R}^{-1}(\hat{c}_s^i) \cdot z_s^i, \\ &\text{Redirection: } \quad \hat{z}_t^i = \mathbf{R}(c_t^i) \cdot z_N^i, \end{aligned} \tag{3}$$

where $i \in \{1, 2\}$, representing gaze directions and head orientations, respectively, and z_N^i denotes the normalized embedding of the corresponding attribute. We feed the original embedding z_s^i and the modified embedding $\hat{z}_t^i$ into the weights-sharing deprojectors ***DP*** to transform these embeddings back to the feature space $\mathcal{F}$ as the residuals. Given these residuals, we implement subtraction and addition operations over f_s as described in Eq. 1 (or Eq. 2) to acquire the edited latent vector $\hat{f}_t$ (or $\hat{f}_{t,d}$), which is sent to a generator for synthesizing redirected face image $\hat{I}_t$. $\hat{I}_t$ should have the same gaze direction and head orientation as I_t.

3.5. Learning Objectives

We supervise the relationship between the generated image $\hat{I}_t$ and the target image I_t with several loss functions: pixel-wise reconstruction loss, LPIPS metric [43] and attributes loss by a task-related pretrained model.

$$\mathcal{L}_{rec}(\hat{I}_t, I_t) = \big|\big|\hat{I}_t - I_t\big|\big|_2, \tag{4}$$

$$\mathcal{L}_{LPIPS}(\hat{I}_t, I_t) = \big|\big|\psi(\hat{I}_t) - \psi(I_t)\big|\big|_2, \tag{5}$$

$$\mathcal{L}_{att}(\hat{I}_t, I_t) = \langle \xi_{hg}(\hat{I}_t), \xi_{hg}(I_t) \rangle, \tag{6}$$

where $\psi(\cdot)$ denotes the perceptual feature extractor [43], $\xi_{hg}(\cdot)$ represents the pretrained HeadGazeNet [44] to estimate the gaze and head pose from images and $\langle u, v \rangle = \arccos \frac{u \cdot v}{||u|| \cdot ||v||}$.

Identity Loss. Identity preservation after redirection is critical for the face editing task. Considering this, we calculate the cosine similarity of the identity-related features between the source image and the redirected image:

$$\mathcal{L}_{ID}(\hat{I}_t, I_s) = 1 - \langle \phi(\hat{I}_t), \phi(I_s) \rangle, \tag{7}$$

where $\phi(\cdot)$ denotes the pretrained ArcFace [11] model.

Label Loss. We have ground truth of gaze directions and head orientations, which can guide the conditions learning from the input latent vectors for the normalization step:

$$\mathcal{L}_{lab}(\hat{c}_s^i, c_s^i) = \langle \hat{c}_s^i, c_s^i \rangle, \quad i \in \{1, 2\}. \tag{8}$$

Embedding Loss. The normalized embeddings only contain the canonical status of the corresponding attribute after the inverse rotation applied to the original estimated embeddings, shown in Fig. 1. Thus the normalized embeddings given a certain attribute across different samples within batch B should be consistent. To reduce the number of possible pairs within a batch, we utilize the first normalized embedding $z_{N,1}^i$ as the basis:

$$\mathcal{L}_{emb} = \frac{1}{B-1} \sum_{j=2}^{B} \langle z_{N,1}^i, z_{N,j}^i \rangle, \quad i \in \{1, 2\}. \tag{9}$$

Layer-wise Weights Loss. This loss is specifically designed for the W^+ space to decide the weights p_i of which layer should contribute more to the aimed attributes editing. Firstly, we calculate the layer-wise estimated conditions $\hat{c}_k^i$ and calculate estimated pseudo labels $\tilde{c}^i$. Secondly, we have layer-wise estimated label errors by $\langle \hat{c}_k^i, \tilde{c}^i \rangle$. Lastly, we calculate the cosine similarity between the reciprocal of label errors and weights of layers as the loss:

$$\mathcal{L}_{prob} = \langle \{p_k\}, \{\frac{1}{\langle \hat{c}_k^i, \tilde{c}^i \rangle}\} \rangle, k \in [1, K], i \in \{1, 2\}, \tag{10}$$

where K is the number of layers for editing.

Full Loss. The combined loss function for supervising the redirection process is:

$$\begin{aligned} \mathcal{L} = &\lambda_r \mathcal{L}_{rec} + \lambda_L \mathcal{L}_{LPIPS} + \lambda_{ID} \mathcal{L}_{ID} + \lambda_a \mathcal{L}_{att} \\ &+ \lambda_l \mathcal{L}_{lab} + \lambda_e \mathcal{L}_{emb} + \lambda_p \mathcal{L}_{prob}, \end{aligned} \tag{11}$$

where $\mathcal{L}_{LPIPS}$ and $\mathcal{L}_{prob}$ are utilized only when the pretrained StyleGAN is used as the generator.

4. Experiments

4.1. datasets

We utilize GazeCapture [24] training subset to train the redirector and assess the performance with its test subset, MPIIFaceGaze and CelebA-HQ [20]. Supplementary material provides more information about these datasets.

4.2. Evaluation Criteria

We follow metrics utilized by ST-ED [44] to evaluate different redirectors' performance.

Redirection Error. We measure the redirection accuracy in image space by a pre-trained ResNet-50 based [16] head pose and gaze estimator ξ'_{hg}, which is unseen during the training. Given the target image I_t and the generated one $\hat{I}_t$ redirected by conditions of I_t, we report the angular error between $\xi'_{hg}(\hat{I}_t)$ and $\xi'_{hg}(I_t)$ as the redirection errors.

Disentanglement Error. We quantify the disentanglement error by the condition's fluctuation range of one attribute when we redirect the other one. The redirection

	Gaze Redir	Head Redir	Gaze Induce	Head Induce	LPIPS
StarGAN † [7]	4.602	3.989	0.755	3.067	0.257
He *et al.* † [17]	4.617	1.392	0.560	3.925	0.223
VecGAN [10]	2.282	0.824	0.401	2.205	**0.197**
ST-ED [44]	2.385	0.800	**0.384**	2.187	0.208
ReDirTrans	**2.163**	**0.753**	0.429	**2.155**	**0.197**

Table 1. Within-dataset quantitative comparison (GazeCapture test subset) between different methods for redirecting head orientations and gaze directions. (Lower is better). **Head (Gaze) Redir** denotes the redirection accuracy in degree between the redirected image and the target image given head orientations (gaze directions). **Head (Gaze) Induce** denotes the errors in degree on gaze (head) when we redirect the head (gaze). † denotes copied results from [44]. Other methods are retrained given previous papers.

	Gaze Redir	Head Redir	Gaze Induce	Head Induce	LPIPS
StarGAN † [7]	4.488	3.031	0.786	2.783	0.260
He *et al.* † [17]	5.092	1.372	0.684	3.411	0.241
VecGAN [10]	2.670	1.242	0.391	1.941	0.207
ST-ED [44]	**2.380**	1.085	**0.371**	**1.782**	0.212
ReDirTrans	**2.380**	**0.985**	0.391	**1.782**	**0.202**

Table 2. Cross-dataset quantitative comparison (MPIIFaceGaze) between different methods for redirecting head orientations and gaze directions. (Lower is better). Notations are the same as them in the Table 1. † denotes copied results from [44]. Other methods are retrained given previous papers.

angle ϵ follows $\mathcal{U}(-0.1\pi, 0.1\pi)$. For example, when we redirect the head pose of the generated image $\hat{I}_t$ by ϵ and generate a new one $\hat{I}'_t$, we calculate the angular error of the estimated gaze directions between $\xi'_{hg}(\hat{I}_t)$ and $\xi'_{hg}(I'_t)$.

LPIPS. LPIPS is able to measure the distortion [43] and image similarity in gaze directions [17] between images, which is applied to evaluate the redirection performance.

4.3. Redirectors in Learnable Latent Space

We compared quantitative performance of different redirectors, which were trained along with the trainable encoder-decoder pair designed by ST-ED on 128 × 128 images with restricted face ranges, given the criteria proposed in Sec. 4.2. Table 1 and Table 2 present within-dataset and cross-dataset performance, respectively. From these tables, we observe that our proposed ReDirTrans achieved more accurate redirection and better LPIPS compared with other state-of-the-art methods by considering the extra embedding space $\mathcal{Z}$ for redirecting embeddings of aimed attributes only and maintaining other attributes including the appearance-related information in the original latent space $\mathcal{F}$. ST-ED [44] projected input latent vectors into nine embeddings including the non-varying embedding z^0. This appearance-related high dimensional embedding z^0 requires more parameters than ReDirTrans during projection. After redirecting the embeddings of aimed attributes, ST-ED deprojected a stack of z^0, redirected embeddings, and rest unvaried embeddings of other attributes back to the feature space for decoding. This projection-deprojection process of non-varying embedding z^0 results in loss of appearance and worse LPIPS, as depicted in Fig. 3. VecGAN [10] was proposed to edit the attributes only within the feature space by addition and subtraction operations. Since there is no projection-deprojection process, given the original latent code, LPIPS performance is better than ST-ED. However, as no extra embedding space was built for the aimed attributes editing, both redirection accuracy and the disentanglement process were affected.

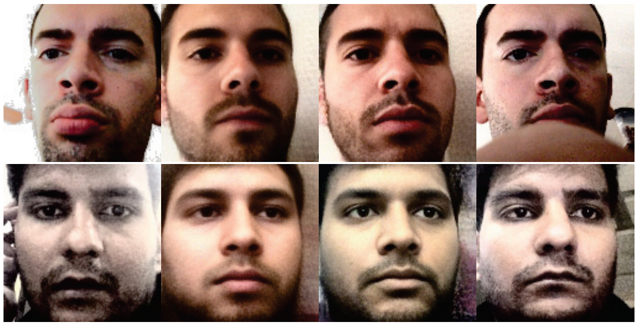
(a) Input (b) ST-ED (c) ReDirTrans (d) Target

Figure 3. Qualitative Comparison of ReDirTrans and ST-ED in GazeCapture. ReDirTrans preserves more facial attributes, such as lip thickness and sharpness of the beard.

4.4. Redirectors in Predefined Latent Space

Except for using the trainable encoder-decoder pair of ST-ED, we also implemented our proposed ReDirTrans within a predefined feature space W^+ to achieve redirection task in full face images with desired resolution. We utilized e4e [35] as the pre-trained encoder, which can transform input images into latent vectors in W^+, and we chose StyleGAN2 [22] as the pre-trained generator to build ReDirTrans-GAN. Fig. 4 shows qualitative comparison between ST-ED and ReDirTrans-GAN in the GazeCapture test subset with providing target images from the same subject. ReDirTrans-GAN successfully redirected gaze directions and head orientations to the status provided by target images while maintaining the same appearance patterns with 1024 × 1024 full face images. Due to the design of ReDirTrans, which maintains unrelated attributes and appearance information in the initial latent space instead of going through the projection-deprojection pro-

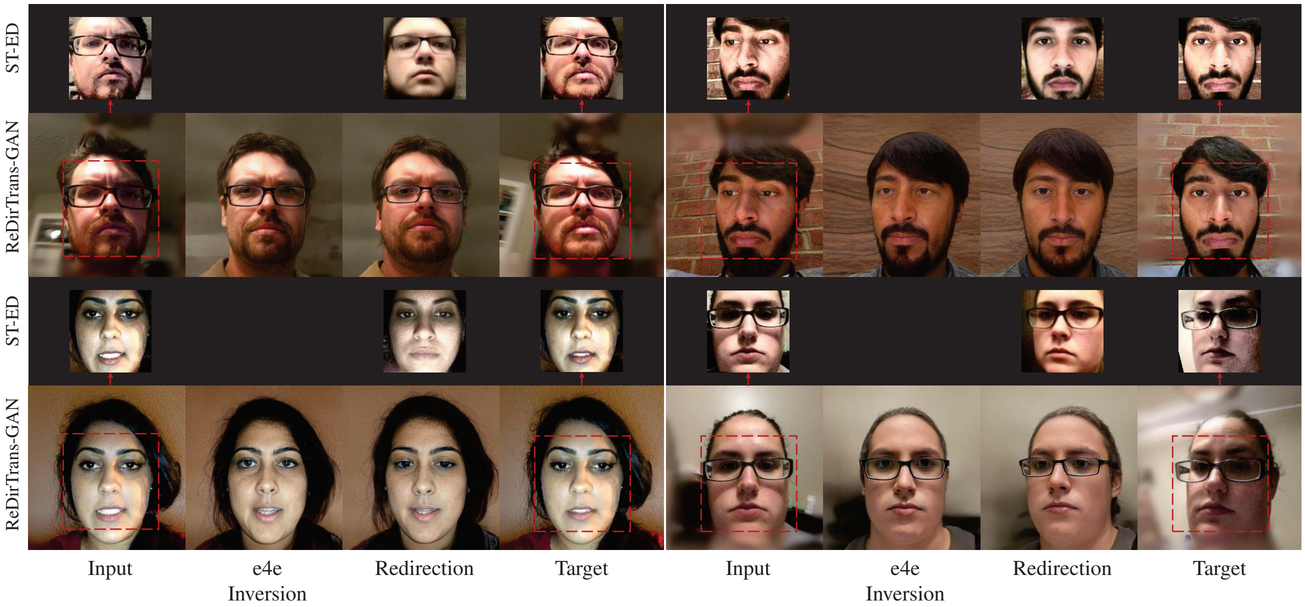

Figure 4. Qualitatively comparisons between ST-ED and ReDirTrans-GAN. Red boxes represent different face covering ranges.

Q%	GazeCapture		MPIIFaceGaze	
	Raw ↓	Aug ↓	Raw ↓	Aug ↓
25	5.875	**5.238**	8.607	**7.096**
50	4.741	**4.506**	6.787	**6.113**
75	4.308	**4.200**	6.165	**5.767**

Table 3. Learning-based gaze estimation errors (in degrees) in GazeCapture and MPIIFaceGaze with or without redirected data augmentation. $Q\%$ represents percent of labeled data in 10,000 images for training ReDirTrans-GAN. 'Raw' or 'Aug' mean training the gaze estimator with real data or with real & redirected data.

cess, ReDirTrans-GAN keeps more facial attributes such as expressions, mustaches, bangs compared with ST-ED. Fig. 5 presents qualitative results with assigned conditions (pitch and yaw of gaze directions and head orientations) in CelebA-HQ [20]. ReDirTrans-GAN can achieve out-of-domain redirection tasks in predefined feature space while maintaining other facial attributes.

4.5. Data Augmentation

To solve data scarcity of the downstream task: learning-based gaze estimation, we utilized redirected samples with assigned gaze directions and head orientations to augment training data. We randomly chose $10,000$ images from the GazeCapture training subset to retrain ReDirTrans-GAN with using only $Q\%$ ground-truth labels of them. The HeadGazeNet $\xi_{hg}(\cdot)$ was also retrained given the same $Q\%$ labeled data and $Q \in \{25, 50, 75\}$. Then we utilized ReDirTrans-GAN to generate redirected samples given provided conditions over $Q\%$ labeled real data and combined the real and redirected data as an augmented dataset with size $2 \times 10,000 \times Q\%$ for training a gaze estimator. Table 3 presented within-dataset and cross-dataset performance and demonstrated consistent improvements for the downstream task given redirected samples as data augmentation.

4.6. Challenge in Predefined Feature Space

One challenge for redirection tasks in predefined feature space comes from inconsistency between input and inverted images, mentioned in Sec. 3.3. We can observe that the existing gaze differences between input and inverted images in Fig. 4. In some cases, the gaze directions are changed after GAN inversion, which means that the encoded latent codes do not necessarily keep the original gaze directions. Thus, instead of using provided gaze directions of input images during the training, we utilized estimated gaze directions from inverted results to correctly normalize the gaze and head pose to the canonical status. This process ensures correctness when further new directions are added, making the training process more consistent.

4.7. Gaze Correction

ReDirTrans can correct gaze directions of inverted results by viewing input images as the target ones. e4e guarantees high editability, which is at the cost of inversion performance [35]. Fig. 6 shows several samples which failed to maintain input images' gaze directions even by the ReStyle encoder [3], which iteratively updates the latent codes given the differences between the input and inverted results. With

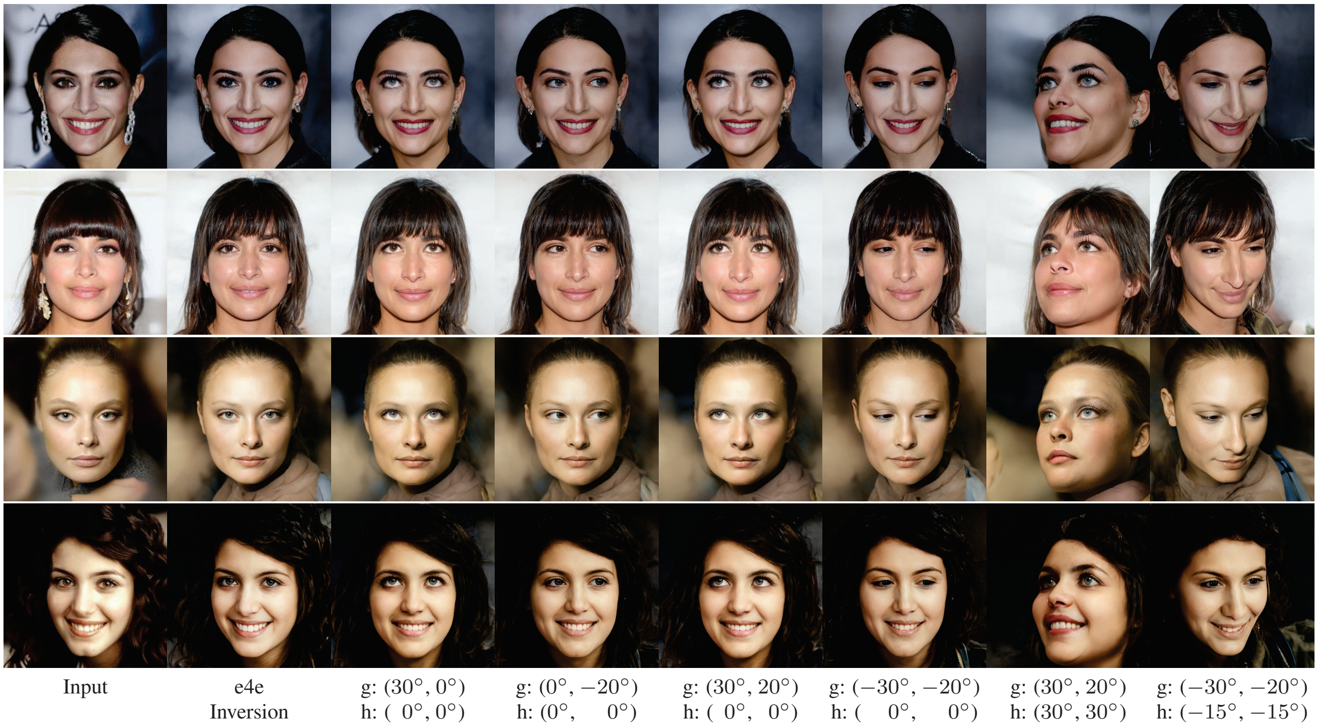
Input
e4e Inversion
g: $(30^\circ, 0^\circ)$ h: $(0^\circ, 0^\circ)$
g: $(0^\circ, -20^\circ)$ h: $(0^\circ, 0^\circ)$
g: $(30^\circ, 20^\circ)$ h: $(0^\circ, 0^\circ)$
g: $(-30^\circ, -20^\circ)$ h: $(0^\circ, 0^\circ)$
g: $(30^\circ, 20^\circ)$ h: $(30^\circ, 30^\circ)$
g: $(-30^\circ, -20^\circ)$ h: $(-15^\circ, -15^\circ)$

Figure 5. Redirection results given assigned (pitch, yaw) conditions of gaze directions (g) and head orientations (h). The first two columns are input and inversion results with e4e [35] and StyleGAN [22]. The following columns are redirected samples with assigned redirection values based on the latent code estimated from e4e.

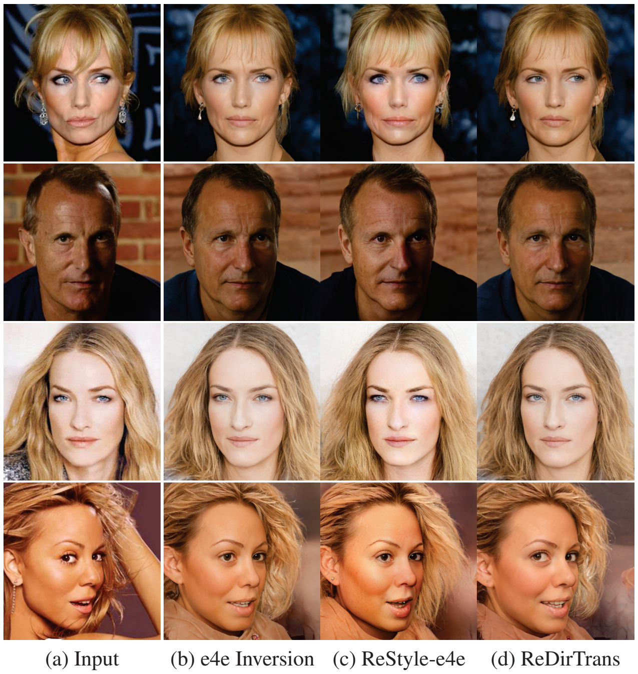
(a) Input (b) e4e Inversion (c) ReStyle-e4e (d) ReDirTrans

Figure 6. Gaze correction in CelebA-HQ by viewing the same image as both the input and target.

ReDirTrans-GAN, we can successfully correct the wrong gaze based on inverted results from e4e.

5. Conclusions

We introduce ReDirTrans, a novel architecture working in either learnable or predefined latent space for high-accuracy redirection of gaze directions and head orientations. ReDirTrans projects input latent vectors into aimed-attribute pseudo labels and embeddings for redirection in an interpretable manner. Both the original and redirected embeddings of aimed attributes are deprojected to the initial latent space for modifying the input latent vectors by subtraction and addition. This pipeline ensures no compression loss to other facial attributes, including appearance information, which essentially reduces effects on the distribution of input latent vectors in initial latent space. Thus we successfully implemented ReDirTrans-GAN in the predefined feature space working with fixed StyleGAN to achieve redirection in high-resolution full-face images, either by assigned values or estimated conditions from target images while maintaining other facial attributes. The redirected samples with assigned conditions can be utilized as data augmentation for further improving learning-based gaze estimation performance. In future work, instead of a pure 2D solution, 3D data can be included for further improvements.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 1, 4

[2] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 1, 2, 4

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 1, 7

[4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022. 1

[5] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 25–32. IEEE, 2014. 1

[6] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. 1

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 6

[8] Min Jin Chong, Wen-Sheng Chu, Abhishek Kumar, and David Forsyth. Retrieve in style: Unsupervised facial feature transfer and retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3887–3896, 2021. 2

[9] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 2

[10] Yusuf Dalva, Said Fahri Altindis, and Aysegul Dundar. Vecgan: Image-to-image translation with interpretable latent directions. *arXiv preprint arXiv:2207.03411*, 2022. 3, 4, 6

[11] 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. 5

[12] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In *Proceedings of the 2017 Chi conference on human factors in computing systems*, pages 1118–1130, 2017. 1

[13] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European conference on computer vision*, pages 311–326. Springer, 2016. 2

[14] René Haas, Stella Graßhof, and Sami S Brandt. Tensor-based emotion editing in the stylegan latent space. *arXiv preprint arXiv:2205.06102*, 2022. 1, 2, 3

[15] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020. 2, 4

[16] 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. 5

[17] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6932–6941, 2019. 1, 2, 6

[18] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. Stressclick: Sensing stress from gaze-click patterns. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1395–1404, 2016. 1

[19] Raphaela E Kaisler and Helmut Leder. Trusting the looks of others: Gaze effects of faces in social settings. *Perception*, 45(8):875–892, 2016. 1

[20] 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. 5, 7

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2

[22] 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, 2, 4, 6, 8

[23] Daniil Kononenko, Yaroslav Ganin, Diana Sungatullina, and Victor Lempitsky. Photorealistic monocular gaze redirection using machine learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2696–2710, 2017. 2

[24] 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. 5

[25] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3835–3844, 2021. 2

[26] 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. 2

[27] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[28] Jiawei Qin, Takuru Shimoyama, and Yusuke Sugano. Learning-by-novel-view-synthesis for full-face appearance-based 3d gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991, 2022. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[30] Julian Schwehr and Volker Willert. Driver's gaze prediction in dynamic automotive scenes. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2017. 1

[31] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 2, 3, 4

[32] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021. 2

[33] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 2

[34] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018. 1

[35] 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, 6, 7, 8

[36] Kang Wang, Rui Zhao, and Qiang Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 440–448, 2018. 2

[37] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. A 3d morphable eye region model for gaze estimation. In *European Conference on Computer Vision*, pages 297–313. Springer, 2016. 2

[38] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 131–138, 2016. 2

[39] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Gazedirector: Fully articulated eye gaze redirection in video. In *Computer Graphics Forum*, volume 37, pages 217–225. Wiley Online Library, 2018. 1

[40] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 1

[41] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019. 1, 2

[42] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020. 1, 2

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 6

[44] 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, 3, 4, 5, 6