Controllable Continuous Gaze Redirection
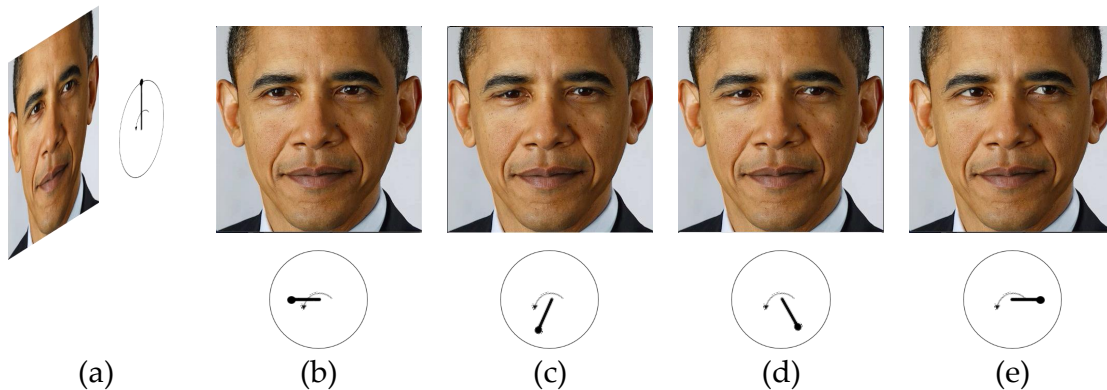
Weihao Xia[1] , Yujiu Yang[1], Jing-Hao Xue[2], Wensen Feng[3]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Department of Statistical Science, University College London
[3]Beijing University of Chemical Technology
xiawh3@outlook.com,yang.yujiu@sz.tsinghua.edu.cn,jinghao.xue@ucl.ac.uk,sanmumuren@126.com

Figure 1: Controllable Continuous Gaze Redirection. (a) A sample image of Barack Obama and the gaze direction space definition; (b-e) Some gaze redirection results from the proposed *interpGaze*. The dots are target directions, and the circles indicate the possible values in the full space of gaze directions.

ABSTRACT

In this work, we present *interpGaze*, a novel framework for controllable gaze redirection that achieves both precise redirection and continuous interpolation. Given two gaze images with different attributes, our goal is to redirect the eye gaze of one person into any gaze direction depicted in the reference image or to generate continuous intermediate results. To accomplish this, we design a model including three cooperative components: an encoder, a controller and a decoder. The encoder maps images into a well-disentangled and hierarchically-organized latent space. The controller adjusts the magnitudes of latent vectors to the desired strength of corresponding attributes by altering a control vector. The decoder converts the desired representations from the attribute space to the image space. To facilitate covering the full space of gaze directions, we introduce a high-quality gaze image dataset with a large range of directions, which also benefits researchers in related areas. Extensive experimental validation and comparisons to several baseline methods show that the proposed *interpGaze* outperforms state-of-the-art methods in terms of image quality and redirection precision.

CCS CONCEPTS

• **Computing methodologies** → **Image processing**; **Computer vision**; **Neural networks**.

KEYWORDS

neural networks, precise gaze redirection, continuous gaze interpolation, continuous image generation

1 INTRODUCTION

Eye contact plays a crucial role in our daily social communication since it conveys important non-verbal cues such as emotion, intention and attention. Gaze redirection is an emerging research topic in computer vision and computer graphics, aiming at manipulating the gaze of a given image to the desired direction with an angle or image as the reference, as illustrated in Figure 1. This task is important in many real-world scenarios. For example, the case that people are not looking at the camera at the same time occurs frequently when taking a group photo. In the video conference, the participants do not have the chance to make direct eye contact due to the location disparity between the video screen and the camera. In both cases, gaze redirection can adjust each person's eye gaze to the same direction or along a single direction to simulate eye

contact. Gaze redirection technique could also alleviate the data insufficiency problem for gaze estimation in the wild, by synthesizing novel samples to augment existing datasets.

Traditional methods typically re-render the entire input region by performing 3D transformations, which requires heavy instrumentation to acquire the depth information [9, 19, 39, 45]. Learning-based gaze redirection approaches [10, 13] has risen in recent years. For example, DeepWarp [10] uses a neural network to warp the input image to the desired direction by predicting the dense flow field. Their method fails to generate photo-realistic images for large redirection angles, especially in the presence of large dis-occlusions, such as large parts of the eyeball being covered by the eyelid in the source image. More importantly, such warping methods cannot generate precise gaze redirection, as it only minimizes the pixel-wise differences between the synthesized and ground-truth images without any geometric regularization.

He *et al.* [13] first introduce Generative Adversarial Network (GAN) to synthesize photo-realistic eye images conditioned on a target gaze direction. Since they incorporate both source gaze image and target gaze vector as input and use a high-quality gaze dataset [29] for training, their method can synthesize images with high quality and redirection precision. However, they do not incorporate head pose together with the vertical and horizontal gaze direction. Head pose is also an important attribute for gaze redirection since the positions of eye balls with the same gaze direction are distinct at different head poses as shown in Figure 2(a). That means they need to train a model for each head pose to precisely control the redirection. Furthermore, although able to produce photo-realistic and precise redirection images, the vector-based methods are prone to the gaze ranges defined in the training dataset and fail to generalize to unseen angles, which hinders the models to cover the full space of gaze directions. Odobez *et al.* [40] propose a self-supervised method to adapt the model pretrained from large amounts of well-aligned synthetic data to real data. Their method produces continuous but perceptually implausible results, due to the limitation of synthetic data as shown in Figure 2(b).

To address the limitations of previous methods, *i.e.*, (1) low-quality generation, (2) low redirection precision and (3) gaze angle limitation, we propose a novel controllable gaze redirection method that can achieve precise redirection and continuous interpolation in one model. More specifically, our method works on both eye image synthesis conditioned on target gaze directions, and continuous intermediate gaze change between two given gaze images. As shown in Figure 3, our model contains an Encoder E, a Controller C and a Decoder G. The Encoder E maps two input gaze images, source image x_s and target image x_t, into a latent feature space with $F_s = E(x_s)$ and $F_t = E(x_t)$. Then the feature difference between F_s and F_t is fed into four branches of the controller C to produce morphing results of two samples $C_v(F_s, F_t)$. The four branches represent head pose, gaze yaw, pitch and miscellaneous attributes respectively. The Decoder G maps the latent features back to the image space. To achieve precise redirection and continuous interpolation, we introduce a gaze vector $v \in [0, 1]^{c \times 1}$ to the controller C. Each element of v corresponds to a mixing indicator for each attribute. Since the head pose, gaze yaw, gaze pitch and miscellaneous attributes have been well-disentangled and hierarchically-organized in the latent space, we can adjust each

attribute by altering the control vector v. For precise gaze redirection, the control vector v selects the features of target directions to generate photo-realistic eye images with the desired head pose and gaze directions. For continuous interpolation, the control vector v interpolates gradually in the latent space of the difference between the source gaze images and the target gaze images. More importantly, benefiting from the flat and smooth latent space, we can generate interpolation sequences in different orders by assigning v, and can generalize to exaggeration of certain attribute by setting corresponding dimension v^k to be a reasonable value greater than 1. The two completely different tasks are unified into the same framework through the proposed feature disentanglement and control mechanism, while avoiding the deficiencies in covering the gaze space of vector-based redirection methods.

In addition, to alleviate the current drawbacks of existing datasets, we collect a large-scale high-quality dataset for gaze redirection and other related tasks, which can also benefit research in related areas. It covers a larger range of eye angles than other popular gaze redirection dataset like Columbia Gaze [29] and is also with diversity on eye shapes, glasses, ages and genders.

Our main contributions can be summarized as follows:

- We present *interpGaze*, a novel framework achieving both precise gaze redirection and continuous gaze interpolation. The two different tasks can be readily controlled by altering a control vector.
- We learn a well-disentangled and hierarchically-organized latent space by decoupling the related gaze attributes and equipping with the efficacy of one-shot and diversity, which makes our method interpretable and complete.
- We contribute a high-quality gaze dataset, which contains a large range of gaze directions and diversity on eye shapes, glasses, ages and genders, to benefit other researchers in related areas.

2 RELATED WORK

Gaze Redirection. Some traditional approaches are based on 3D modeling [2, 36]. They use 3D morphable models to fit both texture and shape of the eye, and re-render the synthesized eyeballs superimposed on the source image. However, these methods make strong model assumptions that may not hold in practice. Therefore, they can not handle images with eyeglasses and other high-variability inter-personal differences. Some others [9, 19, 39, 45] render a scene containing the face of a subject from a given viewpoint to mimic gazing at the camera. These methods require a depth map and synthesize a person-specific image with the redirected gaze by performing 3D transformations.

Learning-based methods have shown remarkable results from using a large dataset labelled with head pose and gaze angles. He *et al.* [13] first introduce Generative Adversarial Network (GAN) to generate photo-realistic eye images while preserving the desired gaze direction. However, the lack of large redirection angles hinders the models to cover the full space of gaze directions. Odobez *et al.* [40] propose a self-supervised method to adapt the model pretrained from large amounts of well-aligned synthetic data to real

(-15P, 0H, 0V)　(0P, 0H, 0V)　(15P, 0H, 0V)　(b)

pitch

(-15P, 15H, 0V)　(0P, 15H, 0V)　(15P, 15H, 0V)

(a)

yaw

(c)
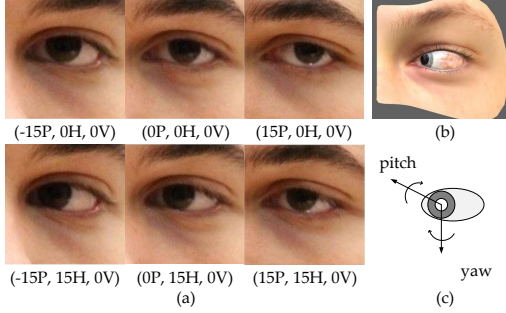
Figure 2: Attribute illustration. (a) Samples from the Columbia Gaze [29]. Typically, gaze-related attributes are head pose (P), horizontal gaze direction (yaw, H) and vertical gaze direction (pitch, V). The directions are illustrated in (c). Number means angle, and negative sign means left hand direction of the subject. Each row of (a) demonstrates the same gaze direction at different head poses, and each column shows the gaze difference at the same head pose. The positions of eye balls with the same gaze direction are distinct at different head poses. (b) A sample from the synthetic UnityEyes [34]. (c) Gaze direction illustration.

data. Their method produces continuous but perceptually implausible results. There are also some face manipulation methods [8, 14] based on image translation or image inpainting to redirect gaze but lack of the capability to precisely control the transformation.

Continuous Image Generation. Recently, there has been some studies on continuous image generation. Most of them can be divided into two categories: (a) Learning explicit or implicit 3D representations and rendering images in different viewpoints. For example, Wood *et al.* [35] fit a parametric eye region model to images and perform gaze redirection by warping eyelids, and compositing eyeballs onto the output in a photo-realistic manner. Chen *et al.* [5] present a transforming auto-encoder in combination with a depth-guided warping procedure that learns to produce geometrically accurate views with fine-grained (*e.g.*, 1° step-size) camera control for both single objects and natural scenes. Olszewski *et al.* [25] propose a *Transformable Bottleneck Networks* to learn the 3D spatial structure from a single image for arbitrary novel-view synthesis. Mildenhall *et al.* [24] represent scenes as *Neural Radiance Fields* for synthesizing novel views from a sparse set of input views. (b) Learning continuous 2D latent features. For example, Pumarola *et al.* [27] introduce a novel GAN model conditioned on a continuous embedding of muscle movements defined by Action Units (AU) annotations, which is able to generate novel facial expressions in a continuum. But their method requires continuous label annotation. Chen *et al.* [7] propose an unsupervised image-to-image translation framework aiming at generating naturally and gradually changing intermediate facial images. Lira *et al.* [22] design a multi-hop mechanism that transforms images gradually between two input domains without any in-between hybrid training images. More recently, Abdal *et al.* [1] propose to directly embed two given images into the latent space of StyleGAN [16], and compute the morphing ω based on a linear interpolation of the obtained vectors ω_1, ω_2

by $\omega = \lambda \cdot \omega_1 + (1 - \lambda) \cdot \omega_2$, where $\lambda \in (0, 1)$ controls the level of mixing of two samples.

3 METHODOLOGY

3.1 Overview

Our goal is to learn a model which can achieve both **precise redirection** and **continuous interpolation**. For an RGB image of an eye patch $x \in \mathbb{R}^{H \times W \times 3}$, we define three primary attributes, *i.e.*, gaze direction $\boldsymbol{d}_g$ and head pose $\boldsymbol{d}_h$, where $\boldsymbol{d}_g = [\phi_g, \theta_g]$, $\phi_g \in \mathbb{R}$ and $\theta_g \in \mathbb{R}$ denote the target yaw and pitch angles, respectively. Given two gaze images x_s and x_t with different attributes, the task is to redirect the eye gaze of one person x_s into any gaze direction x_g depicted in the reference image of another person x_t, or to generate continuous intermediate results x_g between x_s and x_t of the same person.

To achieve this, we design a model including an encoder E, a controller C and a decoder G. The encoder E maps images x_s and x_t into feature space $F_s = E(x_s)$ and $F_t = E(x_t)$. The controller C produces morphing results of two samples $C(F_s, F_t)$. The decoder G maps the desired features back to the image space. Figure 3 provides a full overview of the method. The precise redirection and continuous interpolation can be achieved using the same model by altering a control vector $\boldsymbol{v}$. We elaborate on the three components in more detail below.

3.2 Encoder and Decoder

To make the precise redirection and continuous interpolation feasible, we need the encoder E unfold the natural image manifold to a flat and smooth structure of latent space. We use WGAN-GP [12] to train our model due to its stable performance. The encoder E and the controller C are trained to minimize the Wasserstein distance between real gaze images and generated ones, while a discriminator $\mathcal{D}$ is trained to maximize the distance. It can be formulated as

$$\min_{\mathcal{D}} \mathcal{L}_{GAN_{\mathcal{D}}} = \mathbb{E}_{\mathbb{P}_f}[\mathcal{D}(\hat{F})] - \mathbb{E}_{\mathbb{P}_r}[\mathcal{D}(F)] + \lambda_{gp}\mathcal{L}_{gp},$$
$$\min_{E,C} \mathcal{L}_{GAN_{E,C}} = \mathbb{E}_{\mathbb{P}_r}[\mathcal{D}(F)] - \mathbb{E}_{\mathbb{P}_f}[\mathcal{D}(\hat{F})]. \tag{1}$$

In Equation 1, $F = E(x)$ is the feature extracted from gaze image x by the encoder E; $\hat{F} = C(F_i, F_j)$ is the mixing feature generated by a morphing function of the obtained vectors F_i, F_j; $\mathbb{P}_r$ and $\mathbb{P}_f$ are the distributions of real and fake gaze images respectively; and $\mathcal{L}_{gp}$ is the gradient penalty term [12] used to maintain the Lipschitz continuity of $\mathcal{D}$. The hyperparameter λ_{gp} controls the strength of the gradient penalty, and we use $\lambda_{gp} = 10$ in all experiments. Here the encoder E works together with the controller C to generate target gaze images. We will introduce more details of controller C in the next section.

We additionally incorporate a decoder G to invert features back to images. The decoder G is trained with perceptual loss and reconstruction loss.

The perceptual loss proposed in [15] penalizes the decoder G for generating images that match ground-truth images perceptually. Typically, the perceptual loss is defined by the VGG-16 [28] model pre-trained on ImageNet [18]. The perceptual loss contains two
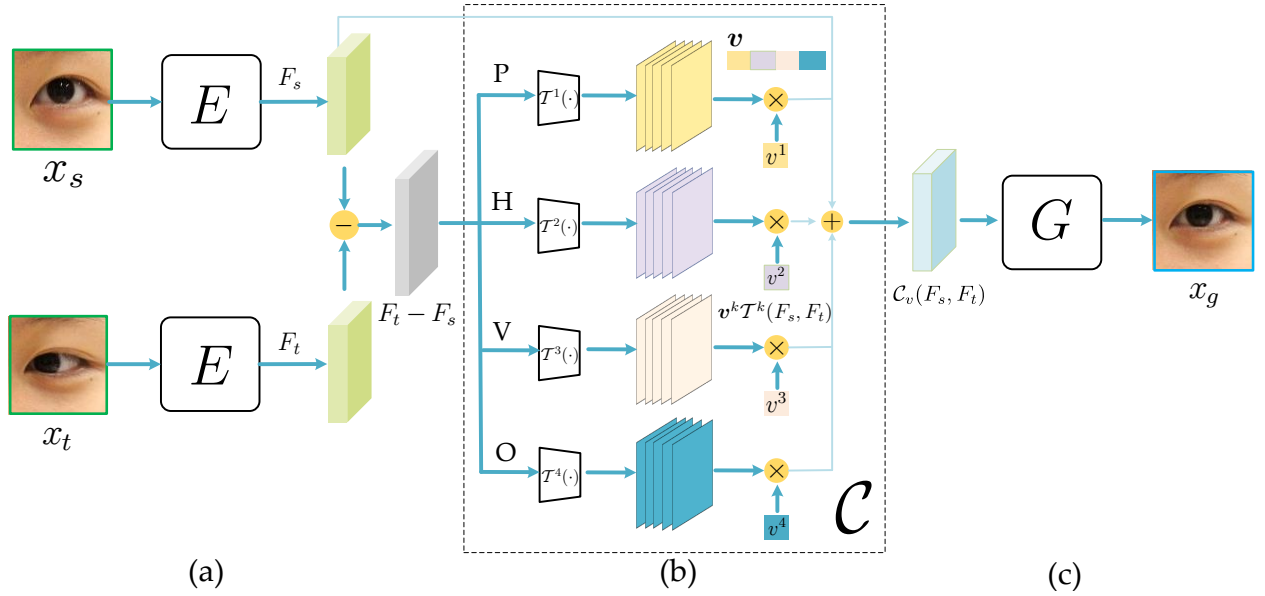
Figure 3: Overview of the proposed *interpGaze*. Our proposed model contains (a) an Encoder E, (b) a Controller C and (c) a Decoder G. The Encoder E maps images x_s and x_t into feature space $F_s = E(x_s)$ and $F_t = E(x_t)$. Then the feature difference is fed into four branches of the controller C to produce morphing results of two samples $C_v(F_s, F_t) = F_s + \sum_{k=1}^{c+1} v^k \mathcal{T}^k(F_t - F_s)$. The abbreviations P, H, V and O are head pose (P), vertical gaze direction (pitch, V), horizontal gaze direction (yaw, H) and miscellaneous attributes. The "O" branch is designed for other secondary attributes like glass, eyebrow, skin color, hair and illumination. The control vector $v \in [0,1]^{(c+1)\times 1}$ adjusts the strength of each attribute, where $c = 3$ in current setting. The Decoder G maps the latent features back to the image space. Please refer to Section 3 for details.
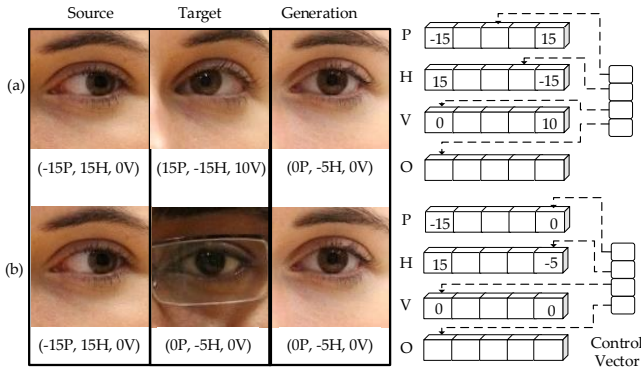

Figure 4: Illustration of Control Mechanism of *interpGaze* for (a) gaze interpolation and (b) gaze redirection. The two completely different tasks are unified into the same framework and can be controlled by altering a control vector. Please refer to Section 3.4 for more details.

terms, the content loss $\mathcal{L}_c$ and style loss $\mathcal{L}_s$, which are defined as

$$\min_{G} \mathcal{L}_c = \mathbb{E}\left(\frac{1}{H_j W_j C_j} \left\| \psi_j(G(F)) - \psi_j(x_s) \right\|^2\right), \quad (2)$$

$$\min_{G} \mathcal{L}_s = \mathbb{E}\left(\sum_{j=1}^{J} \left\| f_j(G(F)) - f_j(x_s) \right\|^2\right), \quad (3)$$

where ψ denote the pre-trained VGG-16 network, $\psi_j(x) \in \mathbb{R}^{H_j \times W_j \times C_j}$ is the activation of j-th layer of ψ, $\mathcal{L}_s$ is the sum of all style losses from the 1-st layer to the J-th layer of the VGG model. $f_j(x)$ denotes the Gram matrix, the entry of which is defined as

$$f_j(x)_{c,c'} = \frac{1}{H_j W_j C_j} \sum_{h}^{H_j} \sum_{w}^{W_j} \psi_j(x)_{h,w,c} \psi_j(x)_{h,w,c'}. \quad (4)$$

The perceptual loss is the sum of content loss and style loss: $\mathcal{L}_p = \mathcal{L}_c + \mathcal{L}_s$. The above two loss terms can force the generated eye patch images to be photo-realistic, and ensure redirection of the gaze directions simultaneously. Following [46], we enforce cycle consistency as the reconstruction term to ensure that personalized features are maintained during the redirection process, which is defined as

$$\min_{E,G} \mathcal{L}_{recon} = \mathbb{E}(\left\| (G(E(x_g))) - x_s \right\|^2). \quad (5)$$

We also introduce a knowledge distillation loss to guide the training of E. Knowledge distillation [3] is widely used to compress CNN models and has recently been utilized for latent space interpolation [7, 31]. However, distilling the dark knowledge from the teacher's output pixels is difficult. Instead, we match the intermediate representations as explored in prior studies [21]. The
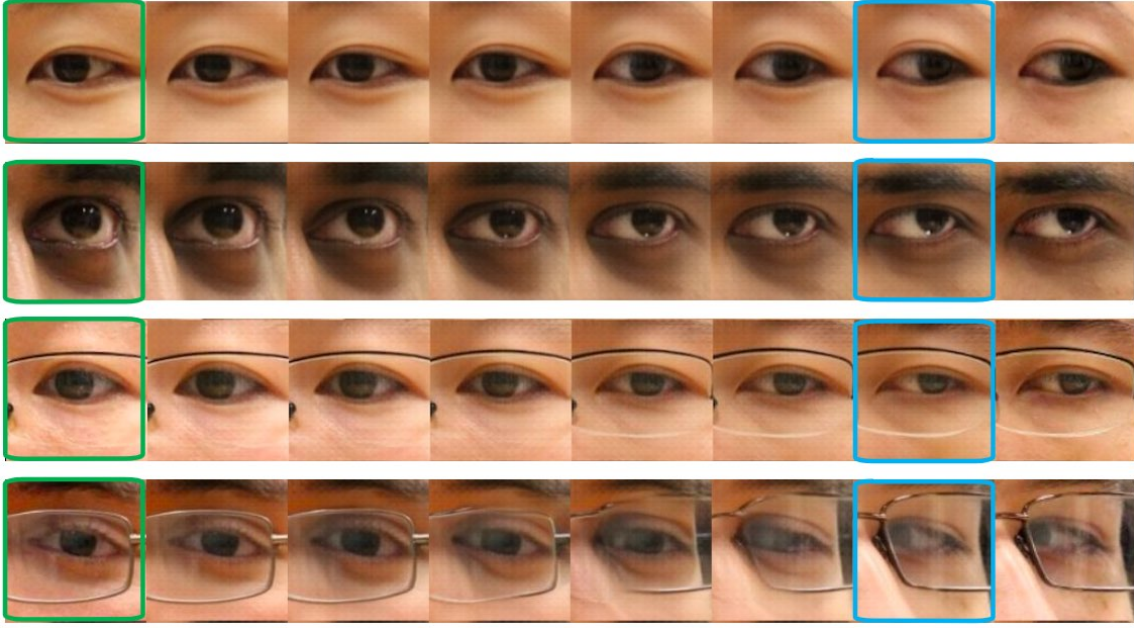
Figure 5: Illustration of gaze interpolation and extrapolation. The four examples are produced by the proposed *interpGaze*. Green rectangles are the start gaze images and blue ones are the ends. Between them are intermediate results of the two given gaze images. The last column is the result of extrapolation. Detailed interpretation can be found in Section 5.3.

intermediate layers allow the student model to acquire more information in addition to outputs, as they contain more channels and provide richer information. The distillation objective can be formalized as

$$\min_{E,F} \mathcal{L}_{distill} = \sum_{t=1}^{T} \|E_t(\mathbf{x}) - F_t(\psi_t(\mathbf{x}))\|_2, \quad (6)$$

where $E_t(\mathbf{x})$ and $\psi_t(\mathbf{x})$ are the intermediate feature activations of the t-th chosen layer in the encoder E and VGG models, and T denotes the number of layers. A 1×1 learnable convolution layer F_t maps the features from the student model to the same number of channels in the features of the teacher model. We jointly optimize E_t and F_t to minimize the distillation loss $\mathcal{L}_{distill}$.

The final objective function of the encoder E and decoder G is

$$\begin{aligned} \mathcal{L}_E &= \lambda_{GAN_E} \mathcal{L}_{GAN_{E,C}} + \lambda_E \mathcal{L}_{recon} + \lambda_{distill} \mathcal{L}_{distill}, \\ \mathcal{L}_G &= \lambda_p \mathcal{L}_p + \lambda_G \mathcal{L}_{recon}, \\ \mathcal{L}_{\mathcal{D}} &= \lambda_{GAN_{\mathcal{D}}} \mathcal{L}_{GAN_{\mathcal{D}}}. \end{aligned} \quad (7)$$

We set all scalars λ_i as 1 in our experiments.

3.3 Controller

The controller C produces morphing results $C(F_s, F_t)$ of two features. Then the decoder G maps the morphed latent features back to the image space. This morphing process typically can be represented as the *Latent Space Interpolation* [3, 4, 6, 32]. Specifically, this process can be done linearly as

$$C(F_i, F_j) = F_i + \alpha(F_j - F_i), \quad (8)$$

where F_i and F_j are two features of real samples, and $\alpha \in [0, 1]$ is a parameter that controls the mixing strength of two samples. The second term $\alpha(F_j - F_i)$ can also be seen as the relative offset or a shifting vector that points from F_i towards F_j. These interpolation methods can connect images of different attributes and produces intermediate results. Nevertheless, they fail to achieve precise control and explain how these attributes are mixed.

The controller C plays two roles: (a) disentangle the representations of these three primary gaze-related attributes and other attributes, and (b) manipulate them precisely.

To obtain disentangled representation for each attribute, we introduce a function $\mathcal{I}(\cdot)$ that maps latent feature F_i to an attribute vector z_i, i.e., $\mathcal{I}(F_i) = z_i$. The translation from the latent space to the attribute space can be performed by the following structure-preserving map (and vice versa), which is analogous to *isomorphism* in algebra [33]:

$$\mathcal{I}(C(F_i, F_j)) = C'(\mathcal{I}(F_i), \mathcal{I}(F_j)), \quad (9)$$

where $C'(z_i, z_j)$ can be viewed as an morphing function that interpolates in the attribute space.

To manipulate these attributes flexibly, we need a control vector $\boldsymbol{v} \in [0, 1]^{c \times 1}$ to adjust the strength of each selected attribute. With both extensions of Equation 8, the revised version of morphing process $C(F_i, F_j)$ can be written as a more flexible formulation $C_{\boldsymbol{v}}(F_i, F_j)$. The linear interpolation defined in Equation 8 is extended to a piece-wise one of

$$C_{\boldsymbol{v}}(F_i, F_j) = F_i + \sum_{k=1}^{c+1} \boldsymbol{v}^k \mathcal{T}^k (F_j - F_i), \quad (10)$$

where $\mathcal{T}^k(\cdot)$ is a learnable mapping function represented by CNN, $\boldsymbol{v}^k$ is the k-th dimension of $\boldsymbol{v}$, $k = 1, \cdots, c, c + 1$, c is the number of the three primary attributes (head pose, gaze yaw or pitch). We also add another branch for other secondary attributes such as glass, skin color, eyebrow and hair.

$\boldsymbol{v}^k$ corresponds to the interpolation strength of the k-th attribute z^k. The k-th attribute changes from sample i to j accordingly as $\boldsymbol{v}^k$ alters from 0 to 1. Thus, interpolation in the latent feature space should correspond to interpolation in the attribute space, if all possible values of z form an attribute space. The relation between the latent space and the attribute space initially defined in Equation 9 can be re-formulated as

$$\mathcal{I}\left(C_{\boldsymbol{v}}\left(F_i, F_j\right)\right) = C_{\boldsymbol{v}}'\left(\mathcal{I}\left(F_i\right), \mathcal{I}\left(F_j\right)\right), \forall \boldsymbol{v} \in [0, 1]^{(c+1) \times 1}. \quad (11)$$

In Equation 11, $C_{\boldsymbol{v}}'(z_i, z_j)$ can be viewed as an interpolation function defined in the attribute space that is controlled by $\boldsymbol{v}$. $C_{\boldsymbol{v}}'(z_i, z_j)$ is defined as

$$C_{\boldsymbol{v}}'\left(z_i, z_j\right) = \left[C_{\boldsymbol{v}}'\left(z_i, z_j\right)^1 \cdots, C_{\boldsymbol{v}}'\left(z_i, z_j\right)^{c+1}\right], \quad (12)$$

where $C_{\boldsymbol{v}}'\left(z_i, z_j\right)^k = z_i^k + \boldsymbol{v}^k\left(z_j^k - z_i^k\right)$. Each dimension $\boldsymbol{v}^c$ sets the interpolation strength of each attribute between the two samples. There is a critical assumption underlying in Equation 11, which is the equivalence of the structure in the latent space and the attribute space defined by operations $C_{\boldsymbol{v}}(\cdot)$ and $C_{\boldsymbol{v}}'(\cdot)$, i.e., isomorphism. The left-hand side of Equation 11 means the attribute values of morphing samples $C_{\boldsymbol{v}}(F_j - F_i)$, and the right side represents attribute values of the two samples. They are expected to be equal since both sides are controlled by the same control vector $\boldsymbol{v}$. To ensure this, we train a network to approximate $C(\cdot)$ and establish a one-to-one mapping from a unique identity element in the latent space to the corresponding one in the attribute space. Specifically, we train the attribute classification network $\mathcal{I}'(\cdot)$ to map the interpolated feature $C_{\boldsymbol{v}}(F_j - F_i) = F_i + \sum_{k=1}^{c+1} \boldsymbol{v}^k \mathcal{T}^k(F_j - F_i)$ to attribute z_i. This can be achieved by minimizing the cross-entropy loss

$$\min_{C_{\boldsymbol{v}}} \mathcal{L}_{C_{isp}} = \mathbb{E}[-C_{\boldsymbol{v}}'(z_i, z_j) \log(\mathcal{I}'(C_{\boldsymbol{v}}(F_i, F_j)))]. \quad (13)$$

During training, we assign uniformly random values to $\boldsymbol{v}$ to cover the whole feasible set.

Besides, similar to the perceptual losses [15], we introduce a loss defined on feature space, which is formulated as

$$\mathcal{L}_{C_t} = \left\| C_{\boldsymbol{v}}\left(F_i, F_j\right) - F_j \right\|^2. \quad (14)$$

To summarize, the overall loss function $\mathcal{L}_C$ of Controller C is

$$\mathcal{L}_C = \lambda_{GAN_C} \mathcal{L}_{GAN_{E,C}} + \lambda_{C_{isp}} \mathcal{L}_{C_{isp}} + \lambda_{C_t} \mathcal{L}_{C_t} \quad (15)$$

where $\mathcal{L}_{GAN_{E,C}}$, $\mathcal{L}_{C_{isp}}$ and $\mathcal{L}_{C_t}$ are defined in Equation 1, Equation 13 and Equation 14, respectively, and λ_{GAN_C}, $\lambda_{C_{isp}}$ and λ_{C_t} are all set to 1 in our experiments.

3.4 Control Mechanism

Figure 4 shows the control mechanism for gaze redirection and interpolation. The abbreviations P, H, V and O are head pose (P), vertical gaze direction (pitch, V), horizontal gaze direction (yaw, H) and miscellaneous attributes, which are also consistent with the four branches in Figure 3.
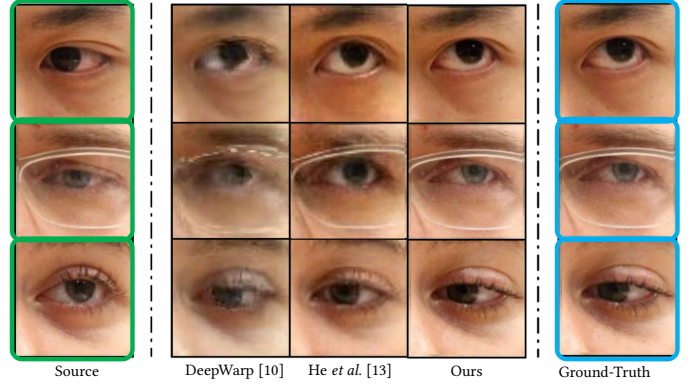

Figure 6: Gaze redirection comparison. Detailed discussion in Section 5.2.

Table 1: Attributes and corresponding labels. The 1st-3rd columns: dimension index in the control vector, name of attributes, corresponding attribute labels.

Dim	Attribute	Labels
1	head pose	$0°, \pm15°, \pm30°$
2	gaze yaw	$0°, \pm5°, \pm10°, \pm15°, \pm25°, \pm35°, \pm45°$
3	gaze pitch	$0°, \pm10°, \pm15°, \pm25°, \pm35°$

For gaze redirection, the inputs are a gaze patch of a certain person as the source image and that of another person (whose appearance may vary considerably) with the desired directions as the target. For gaze interpolation, the inputs are two gaze patches of the same person. We can set the strength and order of the control vector to select features of each attribute and generate the desired results. For example, given $(-15P, 0V, 15H)$ as the source and $(15P, 0V, -15H)$ as the target, we can set the control vector $\boldsymbol{v}_1$ as $[0.5, 0.667, 0, 0]$ to generate a specific intermediate result $(0P, 0V, -5H)$, as illustrated in Figure 4(a). Similarly, we can generate predictable interpolation sequences in different **orders** using different control vectors. If we want to redirect the same source image to the desired direction $(0P, 0V, -5H)$ defined in the reference image, we can set $\boldsymbol{v}_2$ as $[1.0, 1.0, 0, 0]$, which alters the target attribute while keeping others almost intact, as illustrated in Figure 4(b). The appearance of the reference image is allowed to be dramatically different from the source. This means that our method can be extended into **one-shot** [23] inference, using the synthetic image from UnityEyes [34] or SynthesEyes [37], which covers the full space of gaze direction, as the reference to produce an image with the unseen target direction.

The "O" branch in Figure 4 represents other attributes like glass, eyebrow, skin color and hair as aforementioned. Both the last elements in $\boldsymbol{v}_1^4$ and $\boldsymbol{v}_2^4$ are set as 0, which means that these secondary attributes remain unchanged during the generation. But we can also set $\boldsymbol{v}_1^4$ as any value in $[0, 1]$ to generate **diverse** results, which makes our method multi-modal [20, 38, 47]. On the one hand, the "0" branches describe the common attributes of the same person when performing interpolation. On the other hand, the appearance of the
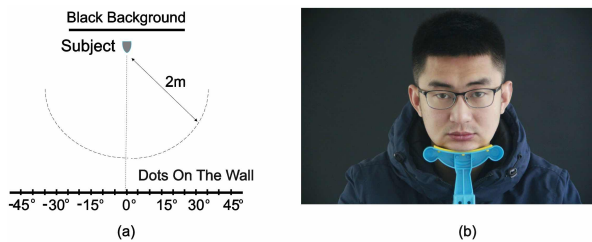
Figure 7: Data Collection Setup (a) and an example (b). For more details, please see Section 4.

two input images may be slightly different (not the attributes of the subject in the picture), which is mostly caused by the light and shade during photo capturing. This also indicates that our method can be extended to integrate more attributes like illumination and exposure.

Through the proposed feature disentanglement and control mechanism, we have unified two completely different tasks into the same framework, while avoiding the deficiencies in covering the gaze space of vector-based redirection methods. Notice that we use relative offset of each attribute of two inputs in practice, the illustration in Figure 4 is simplified for easy understanding.

4 DATASET

4.1 Why Do We Need A New Gaze Dataset?

Unfolding the natural image manifold to a flat and smooth latent space is the key for our method to perform precise redirection and continuous interpolation, including the more challenging one-shot and extrapolation. In addition to designing the network structure to allow the encoder to obtain a smooth and flat latent space, a dataset with a relatively larger range of directions is also crucial for the model to learn such latent space. Furthermore, such a high-quality gaze dataset can also benefit other researchers in related areas.

4.2 Dataset Collection

We collect images in the same way as [29]. As shown in Figure 7, subjects were seated in a fixed location in front of a black background, and a 13×9 grid of dots was attached to a wall in front of them. The dots were placed by the angle-distance relation. For each subject and head pose, we took images of the subject gazing at each dot of the pose's corresponding grid of dots. Then, we collect a high-quality gaze dataset that contains $29, 250$ images of 50 subjects with the resolution of $5, 184 \times 3, 456$. We use resized eye patch images following the same procedure as in [13] for our experiments. We have three gaze-related attributes with certain labels, *i.e.*, head pose, gaze yaw and pitch direction, as shown in Table 1. We randomly select 85 subjects from the Columbia Gaze [29] and ours for training, and use the others for testing. For fair comparison on gaze redirection, we also train our model only with Columbia Gaze [29].

4.3 Dataset Analysis

There are several important properties that a high-quality gaze redirection dataset should own: realness, high-resolution, collected in the constrained environment, with precise annotation and a large range of directions. We compare a range of RGB gaze datasets in Table 2. Some publicly available gaze datasets, such as MPIIGaze [43], Gaze Capture [17], UT Multi-View [30] or CMU Multi-Pie [11], only provide low-resolution images and would therefore introduce an unwanted bias towards low-quality images. Those [17, 43] proposed for gaze estimation are collected under unconstrained environmental conditions and vary in background and illumination, which means these datasets are not suitable for high-quality gaze redirection. Some recent learning-based gaze redirection studies either learn from synthetic data [40] (*e.g.*, SynthesEyes [37], UnityEyes [34]), or use limited high-quality dataset [13] (*e.g.*, Columbia Gaze [29]). The synthetic data *e.g.*, SynthesEyes [37] and UnityEyes [34], covers the full space of gaze direction but methods trained on them are not as photo-realistic as those trained on high-quality real images. The Columbia Gaze dataset [29] is a high-resolution, publicly available human gaze dataset collected from 56 subjects in a constrained environment. The head poses of each subject are discrete values in the set $[0°, \pm15°, \pm30°]$. For each head pose, there are 21 gaze directions, which are the combinations of three pitch angles $[0°, \pm10°]$, and seven yaw angles $[0°, \pm5°, \pm10°, \pm15°]$. However, it does not cover an enough range of gaze direction. According to the Driver and Vehicle Licensing Agency (DVLA) in the United Kingdom, the normal range of the human eye is about $120°$ of horizontal field and $40°$ of vertical field. In our collection procedure, we found that it is about $45°$ horizontally and $35°$ vertically if focused accurately on a dot. Out of this range, the obtained gaze images would look like almost the same in a certain direction.

5 EXPERIMENTS

5.1 Metrics

It remains an open problem to effectively evaluate the quality of generated images in image generation tasks. Gaze redirection models are required to be precise in redirecting and to produce photo-realistic and consistent images. Correspondingly, the evaluation metrics need to be able to assess these aspects. Similarly to He *et al.* [13], we use the mean squared error (MSE), Learned Perceptual Image Patch Similarity (LPIPS) [42] and Gaze Estimation Error [26], as the metrics to measure the similarity, perception quality and gaze direction error, respectively.

5.2 Gaze Redirection

Baseline Model. We adopt DeepWarp [10] and a recent GAN-based method [13] as our baselines. We did not compare with recent GazeAdaptation [41] and GazeDirector [36] since their implementations are not publicly available.

Qualitative Evaluation. Figure 6 shows the generated gaze images examples. The reference image for certain gaze direction is randomly selected. Although both DeepWarp [10] and He *et al.* [13] are capable of redirecting the gaze, the generated images have several obvious defects. For example, DeepWarp [10] prodeces blurry textures and boundaries such as skin and eyebrows, and the shapes of certain areas, such as the edges of eyelid, iris and eyeglasses, are distorted; the generated gaze images of He *et al.* [13] are more

Table 2: Comparison of some popular gaze datasets.

Dataset	Real	High Res	Constrained	Annotation Type	Num. Image	Head Pose	Gaze Range
CMU Multi-Pie [11]	✓	✗	✓	Facial landmarks	755,370	✓	Small
Gaze Capture [17]	✓	✗	✗	2D position on screen	>2.5M	✗	Small
SynthesEyes [37]	✗	✓	✗	Gaze vector	11,382	✓	Full
UnityEyes [34]	✗	✓	✗	Gaze vector	1,000,000	✓	Full
MPII Gaze [43, 44]	✓	✗	✗	Gaze vector	213,659	✓	Small
UT Multi-View [30]	✓	✗	✓	Gaze vector	1,152,000	✓	Large
Columbia [29]	✓	✓	✓	Gaze vector	5,880	✓	Medium
Ours	✓	✓	✓	Gaze vector	29,250	✓	Large

faithful to the input images, but suffer from unnatural pupil when redirecting to extreme angles as shown in the last row of Figure 6. In contrast, our method achieves better performance on visual plausibility and redirection precision.

Quantitative Evaluation. Quantitative evaluation is summarized in Table 3. We can see that our method achieves the lowest LPIPS score [42] at every correction angle, which indicates that our method can generate gaze images that are perceptually more similar to the ground-truth images. This observation is consistent with the qualitative evaluation.

User Study. We ask 21 users to pick the gaze image that looks more realistic than the other. The generated gaze images are split into three groups as in [13], $[4.9°, 15.0°]$, $(15.0°, 25.0°]$, $(25.0°, 35.9°]$. In each group, we randomly choose 19 images generated by both methods with the same input image and gaze direction. For ours, we randomly choose a gaze image with the desired direction as the reference image. Three images are displayed alongside and shuffled randomly. The results are shown in Table 4.

5.3 Gaze Interpolation

The results of continuous gaze interpolation is shown in Figure 5, where x_s (green rectangles) and x_t (blue rectangles) are two gaze images randomly sampled from the same person, and moving from source image x_s toward target image x_t in the latent space gradually produces continuous realistic images x_g. Given different control vectors v, we can select the strength of intermediate results. For example, the first and third rows change only gaze directions with constant head pose, while the second and forth rows perform the opposite. It can be seen that other attributes like eyebrow, glass, hair and skin color are well-preserved in the redirected gaze images, which means that our model works consistently well in generating person-specific gaze images. We also gradually increase the strength of the last dimension of v to produce multi-modal results, as shown in the third row. Furthermore, our method can also perform extrapolation, as demonstrated in the last column of Figure 5. This indicates that the encoder has unfolded the natural image manifold, leading to a well-learned flat and smooth latent space that allows precise control including redirection, interpolation and extrapolation.

Table 3: Comparison with the state-of-the-arts in terms of the MSE, LPIPS and Gaze Estimation Error of different direction groups. ↓ means the lower the better.

Method	MSE ↓	LPIPS↓	Gaze Error↓
Deepwarp [10]	126.71	0.083	15.3°
He *et al.* [13]	72.15	0.066	8.7°
Ours	**56.90**	**0.037**	**6.3°**

Table 4: Voting results from the user study comparing Deep-Warp [10] and He *et al.* [13] with our method. Each row sums up to 100%.

Group	DeepWarp [10]	He *et al.* [13]	Ours
$[4.9°, 15.0°]$	4.7%	28.6%	**66.7%**
$(15.0°, 25.0°]$	14.3%	33.3%	**52.4%**
$(3.2°, 29.8°]$	9.5%	42.9%	**47.6%**

6 CONCLUSION

In this paper, we have introduced *interpGaze*, a framework for controllable gaze redirection that can easily achieve both precise redirection and continuous interpolation, and even extrapolation by altering a control vector. Through the proposed feature disentanglement and control mechanism, we have unified two completely different tasks into the same framework, while avoiding the deficiencies in covering the gaze space of vector-based redirection methods. With our elaborate network design and our new dataset of large range gaze directions, our method can cover a large space of directions for gaze redirection and interpolation.

7 ACKNOWLEDGEMENT

REFERENCES

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? *International Conference on Computer Vision* (2019).

[2] Michael Banf and Volker Blanz. 2009. Example-based rendering of eye movements. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 659–666.

[3] Yoshua Bengio, Grégoire Mesnil, Yann N. Dauphin, and Salah Rifai. 2013. Better Mixing via Deep Representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013.* 552–560.

[4] David Berthelot, Colin Raffel, Aurko Roy, and Ian J. Goodfellow. 2019. Understanding and Improving Interpolation in Autoencoders via an Adversarial Regularizer. In *Proceedings of International Conference on Learning Representations*.

[5] Xu Chen, Jie Song, and Otmar Hilliges. 2019. Monocular Neural Image Based Rendering with Continuous View Control. In *International Conference on Computer Vision*.

[6] Ying-Cong Chen, Huaijia Lin, Michelle Shu, Ruiyu Li, Xin Tao, Xiaoyong Shen, Yangang Ye, and Jiaya Jia. 2018. Facelet-Bank for Fast Portrait Manipulation. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*. 3541–3549.

[7] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. 2019. Homomorphic Latent Space Interpolation for Unpaired Image-to-image Translation. In *CVPR*.

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8789–8797.

[9] Antonio Criminisi, Jamie Shotton, Andrew Blake, and Philip HS Torr. 2003. Gaze Manipulation for One-to-one Teleconferencing.. In *ICCV*, Vol. 3. 13–16.

[10] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. 2016. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European Conference on Computer Vision*. Springer, 311–326.

[11] R Gross, I Matthews, J Cohn, T Kanade, and S Baker. 2010. Multi-PIE. In *Image and Vision Computing*. 807–813.

[12] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*. 5767–5777.

[13] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. 2019. Photo-realistic Monocular Gaze Redirection using Generative Adversarial Networks. *International Conference on Computer Vision* (2019).

[14] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Trans. Graph.* 36, 4 (2017), 107:1–107:14.

[15] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[16] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*. 4401–4410.

[17] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *CVPR*. 2176–2184.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[19] Claudia Kuster, Tiberiu Popa, Jean-Charles Bazin, Craig Gotsman, and Markus Gross. 2012. Gaze correction for home video conferencing. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 174.

[20] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. 2018. Diverse Image-to-Image Translation via Disentangled Representations. In *ECCV*.

[21] Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han. 2020. GAN Compression: Efficient Architectures for Interactive Conditional GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[22] Wallace P. Lira, Johannes Merz, Daniel Ritchie, Daniel Cohen-Or, and Hao (Richard) Zhang. 2020. GANHopper: Multi-Hop GAN for Unsupervised Image-to-Image Translation. *CoRR* abs/2002.10102 (2020).

[23] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. 2019. Few-Shot Unsupervised Image-to-Image Translation. In *IEEE International Conference on Computer Vision (ICCV)*.

[24] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. 2019. Transformable Bottleneck Networks. In *The IEEE International Conference on Computer Vision (ICCV)*.

[26] Seonwook Park, Adrian Spurr, and Otmar Hilliges. 2018. Deep pictorial gaze estimation. In *ECCV*. 721–738.

[27] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. 2019. GANimation: One-Shot Anatomically Consistent Facial Animation. (2019).

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

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

[30] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2014. Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2018. It Takes (Only) Two: Adversarial Generator-Encoder Networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*. 1250–1257.

[32] Paul Upchurch, Jacob R. Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Q. Weinberger. 2017. Deep Feature Interpolation for Image Content Changes. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*. 6090–6099.

[33] E. K. Wong. 1992. Model matching in robot vision by subgraph isomorphism. *Pattern Recognit.* 25, 3 (1992), 287–303.

[34] Erroll Wood, Tadas Baltrusaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2016. Learning an appearance-based gaze estimator from one million synthesised images. In *ACM Symposium on Eye Tracking Research & Applications, ETRA*. 131–138.

[35] Erroll Wood, Tadas Baltrusaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2018. GazeDirector: Fully Articulated Eye Gaze Redirection in Video. *Comput. Graph. Forum* 37, 2 (2018), 217–225.

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

[37] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation. In *International Conference on Computer Vision*. 3756–3764.

[38] Weihao Xia, Yujiu Yang, and Jing-Hao Xue. 2019. Unsupervised Multi-Domain Multimodal Image-to-Image Translation with Explicit Domain-Constrained Disentanglement. (2019). http://arxiv.org/abs/1911.00622

[39] Ruigang Yang and Zhengyou Zhang. 2002. Eye gaze correction with stereovision for video-teleconferencing. In *European Conference on Computer Vision*. Springer, 479–494.

[40] Yu Yu, Gang Liu, and Jean-Marc Odobez. 2019. Improving Few-Shot User-Specific Gaze Adaptation via Gaze Redirection Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 11937–11946.

[41] Yu Yu, Gang Liu, and Jean-Marc Odobez. 2019. Improving Few-Shot User-Specific Gaze Adaptation via Gaze Redirection Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11937–11946.

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

[43] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based Gaze Estimation in the Wild. In *CVPR*. 4511–4520.

[44] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2299–2308.

[45] Jiejie Zhu, Ruigang Yang, and Xueqing Xiang. 2011. Eye contact in video conference via fusion of time-of-flight depth sensor and stereo. *3D Research* 2, 3 (2011), 5.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.

[47] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*.