# Coarse-to-Fine Gaze Redirection with Numerical and Pictorial Guidance

Jingjing Chen[1], Jichao Zhang[2], Enver Sangineto[2], Tao Chen[3]*, Jiayuan Fan[4], Nicu Sebe[2,5]
[1]Zhejiang University, [2]University of Trento
[3]School of Information Science and Technology, Fudan University
[4]Academy for Engineering and Technology, Fudan University, [5]Huawei Research Ireland

## Abstract

*Gaze redirection aims at manipulating the gaze of a given face image with respect to a desired direction (i.e., a reference angle) and it can be applied to many real life scenarios, such as video-conferencing or taking group photos. However, previous work on this topic mainly suffers of two limitations: (1) Low-quality image generation and (2) Low redirection precision. In this paper, we propose to alleviate these problems by means of a novel gaze redirection framework which exploits both a numerical and a pictorial direction guidance, jointly with a coarse-to-fine learning strategy. Specifically, the coarse branch learns the spatial transformation which warps input image according to desired gaze. On the other hand, the fine-grained branch consists of a generator network with conditional residual image learning and a multi-task discriminator. This second branch reduces the gap between the previously warped image and the ground-truth image and recovers finer texture details. Moreover, we propose a numerical and pictorial guidance module (NPG) which uses a pictorial gazemap description and numerical angles as an extra guide to further improve the precision of gaze redirection. Extensive experiments on a benchmark dataset show that the proposed method outperforms the state-of-the-art approaches in terms of both image quality and redirection precision. The code is available at* `https://github.com/jingjingchen777/CFGR`

## 1. Introduction

Gaze redirection is a new research topic in computer vision and computer graphics and its goal is to manipulate the eye region of an input image, by changing the gaze according to a reference angle. This task is important in many real-world scenarios. For example, when taking a group photo, it rarely happens that everyone is simultaneously looking at the camera, and adjusting each person's gaze with respect to the same direction (e.g., the camera direction) can make the photo look better and user acceptable. In another application scenario, when talking in a video conferencing system, eye contact is important as it can express attentiveness and confidence. However, due to the location disparity between the video screen and the camera, the participants do not have direct eye contact. Additionally, gaze redirection tasks can be applied to improve few-shot gaze estimation [27, 28] and domain transfer [11].

Traditional methods are based on a 3D model which re-renders entire input region [1, 22]. These methods suffer from two major problems: (1) it is not easy to render the entire input region and (2) they require an heavy instrumentation. Another type of gaze redirection is based on machine learning for image re-synthesis, such as DeepWarp [6] or PRGAN [8]. DeepWarp [6] employs a neural network to predict the dense flow field which is used to warp the input image with respect to the gaze redirection. However, this method cannot generate perceptually plausible samples, as only using the pixel-wise differences between the synthesized and ground truth images is insufficient. PRGAN [8] proposes a GAN-based autoencoder with a cycle consistent loss for monocular gaze redirection and it can synthesize samples with high quality and redirection precision. However, its single-stage learning causes the corresponding appearance to look asymmetric. Overall, the previous results are still far from the requirements imposed by many application scenarios.

In this paper, we propose a coarse-to-fine strategy and we combine flow learning with adversarial learning to produce higher quality and more precise redirection results. As shown in Fig. 1, our model consists of three main parts. The first one is a coarse-grained model which is an encoder-decoder architecture with flow learning and models the eye spatial transformation. Specifically, this network is fed with source images and with the angle-difference vector between target and source. Second, in order to refine the warped results, we propose to use a conditional architecture, in which the generator learns the residual image between the warped output and the ground truth. The goal of the generator is to

*Tao Chen is the corresponding author.

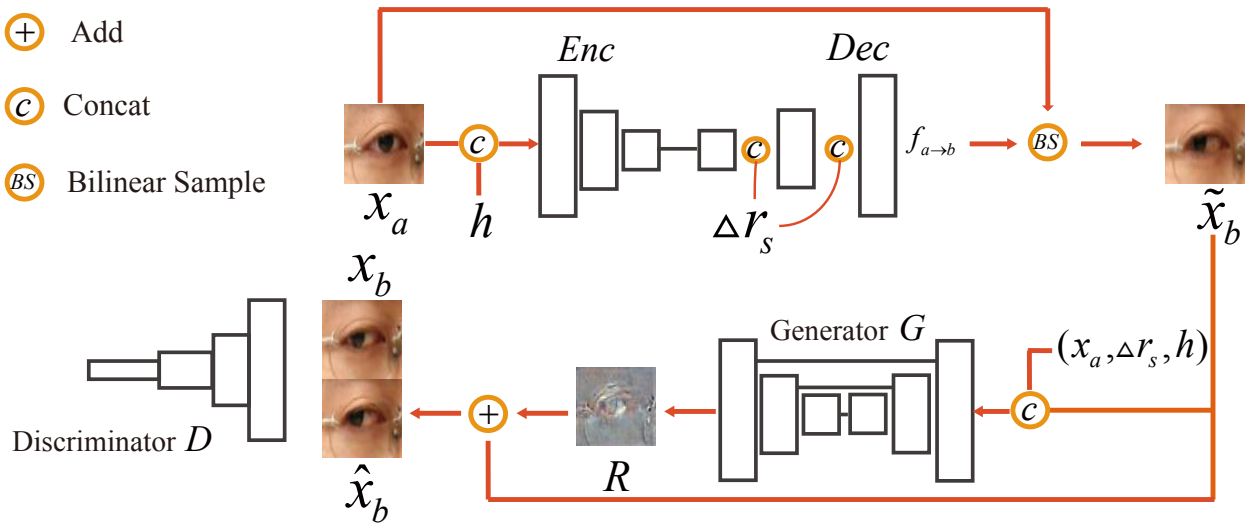


Figure 1. The pipeline of the proposed gaze redirection approach. The upper branch outputs a coarse-grained result $\tilde{x}_b$. The encoder $Enc$ takes as input the eye region $x_a$ and the head pose $h$, while the decoder $Dec$ takes as input the encoder latent code and $\triangle r_s$ (provided by the NPG module, not shown in the figure). The lower branch outputs fine-grained, final results. The generator $G$ outputs the residual image $R$, which is added to $\tilde{x}_b$. The refined results $\hat{x}_b$ and the ground truth $x_b$ are fed to the discriminator $D$.

reduce possible artifacts in the warped texture and the distortions in the eye shape. Finally, a discriminator network with gaze regression learning is used to ensure that the refined results have the same distribution and the same gaze angles as the ground truth. Additionally, we propose an NPG module which integrates the pictorial gazemap representation with numerical angles to guide the synthesis process. The intuitive idea is that the gazemap pictorial representation can provide additional spatial and semantic information of the target angle (shown in Fig. 2).

The main contributions of our work are:

1. We propose a coarse-to-fine eye gaze redirection model which combines flow learning and adversarial learning.

2. We propose an NPG module which integrates the pictorial gazemap with numerical angles to guide the generation process.

3. We present a comprehensive experimental evaluation demonstrating the superiority of our approach in terms of both image quality of the eye region and angle redirection precision.

## 2. Related Work

**Facial Attribute Manipulation**, an interesting multi-domain image-to-image translation problem, aims at modifying the semantic content of a facial image according to a specified attribute, while keeping other irrelevant regions unchanged. Most works [3, 30, 14, 17, 18, 25, 2, 31, 9, 34, 15, 21] are based on GANs and have achieved impressive facial attribute manipulation results. However, these methods tend to learn the style or the texture translation and are not good in obtaining high-quality, natural geometry translations. To alleviate this problem, Yin, et al. [26] proposed a geometry-aware flow which is learned using a geometry guidance obtained by facial landmarks. Wu, et al. [24] also exploits the flow field to perform spontaneous motion, achieving higher quality facial attribute manipulation. Eye gaze redirection can be considered as a specific type of facial attribute manipulation. To the best of our knowledge, our model is the first combining flow learning and adversarial learning for gaze redirection.

**Gaze Redirection.** Traditional methods are based on a 3D model which re-renders the entire input region. Banf and Blanz [1] use an example-based approach to deform the eyelids and slides the iris across the model surface with texture-coordinate interpolation. GazeDirector [22] models the eye region in 3D to recover the shape, the pose and the appearance of the eye. Then, it feeds an acquired dense flow field corresponding to the eyelid motion to the input image to warp the eyelids. Finally, the redirected eyeball is rendered into the output image.

Recently, machine learning based methods have shown remarkable results using a large training set labelled with eye angles and head pose information. Kononenko and Lempitsky [13] use random forests as the supervised learners to predict the eye flow vector for gaze correction. Ganin et al. [6] use a deep convolution network with a coarse-to-fine warping operation to generate redirection results. However, these warping methods based on pixel-wise differences between the synthesized and ground-truth images, have difficulties in generating photo-realistic images and

they fail in the presence of large redirection angles, due to dis-occlusion problems. Recently, PRGAN [8] adopted a GAN-based model with a cycle-consistent loss for the gaze redirection task and succeeded in generating better quality results. However these results are still far from being satisfactory. To remedy this, Zhang, et al. [29] developed a dual inpainting module, to achieve high-quality gaze redirection in the wild by interpolating the angle representation. However, also this method fails to redirect gaze with arbitrary angles.

Compared to the previous methods, our approach exploits a coarse-to-fine learning process and it learns the flow field for the spatial transformation. This is combined with adversarial learning to recover the finer texture details. Moreover, we are the first to propose utilizing the gaze map (i.e., the pictorial gaze representation) as an input to provide extra spatial and semantic information for gaze redirection. Empirically, we found that this is beneficial in order to improve the redirection precision.

## 3. Method

The pipeline of the proposed method is shown in Fig. 1. It is mainly split into two learning stages. In the coarse learning stage, an encoder-decoder architecture is proposed to generate coarse-grained results by learning the flow field necessary to warp the input image. On the other hand, the fine learning stage is based on a multi-task conditional GAN, in which a generator with conditional residual-image learning refines the coarse output and recovers finer texture details, eliminating the distortion in the eye geometry. Moreover, we propose an NPG module to guide both the coarse and the fine process (see Fig 2). Before introducing the details, we first clarify the adopted notations.

• Two angle domains: source domain $A$ and target domain $B$. Note that paired samples exist in the two domains.

• $(x_a, r_a) \in A$ indicates the input eye image $x_a \in R^{m\times n\times c}$ from domain A and its corresponding angle pair $r_a \in R^2$, representing the eyeball pitch and yaw $[\theta, \phi]$. $(x_b, r_b) \in B$ are defined similarly. With $m, n, c$ we indicate, respectively, the width, the height and the channel number of the eye image. $x_a$ and $x_b$ are paired samples with different labeled angles. Our model learns the gaze redirection from $A$ to $B$.

• $\triangle r$ denotes the angle vector difference between an input sample in $A$ and the corresponding target sample in $B$.

• $S \in R^{m\times n\times 2}$ denotes the two channel gazemap pictorial representation (the eyeball and the iris), which is generated from an angle pair $r$. $S = F_s(r)$, where $F_s$ is a straightforward graphics tool used to generate pictorial representations of the eyeball and the iris, as shown in Fig. 2. Note that each instance $x_a$ with same angle from domain $A$ has the same $S$.

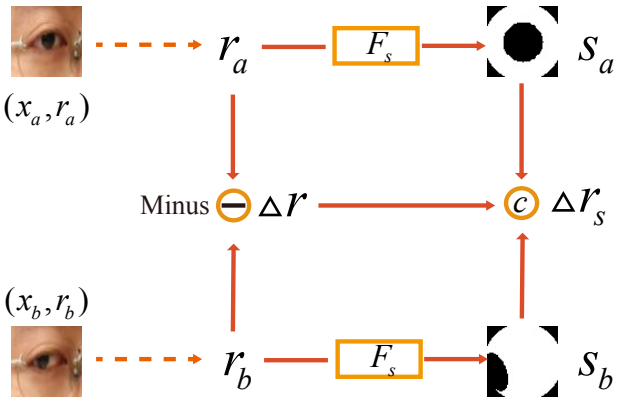


Figure 2. A scheme of the Numerical and Pictorial Guidance Module (NPG). The angle difference vector $\triangle r$ is concatenated with two gazemaps $[S_a, S_b]$ to get $\triangle r_s$. $S_a$ and $S_b$ correspond to the input angles $r_a$ and the target angles $r_b$, respectively. Note that the gazemap has a dimension different from the numeric angle $r$, thus a scale normalization is necessary. $F_s$ is the graphic tool producing the pictorial representation.

### 3.1. Flow-Field Learning for Coarse-Grained Gaze Redirection

To redirect $x_a$ with an angle pair $r_a$ from domain $A$ to domain $B$, our encoder $Enc$ takes both $x_a$ and the corresponding head pose $h$ as inputs. Then, the decoder $Dec$ generates a coarse-grained output using both the encoded code and $\triangle r_s$ (provided by the NPG, see later). As shown in Fig. 1, $\triangle r_s$ is concatenated into different scales of $Dec$ to strengthen the guided ability of the conditional information. This can be formulated as follows:

$$f_{a\to b} = Dec(Enc(x_a, h), \triangle r_s), \tag{1}$$

where $f_{a\to b}$ is the learned flow field from $x_a$ to $x_b$. Similarly to DeepWarp [6], we generate the flow field to warp the input image. In more details, the last convolutional layer of $Dec$ produces a dense flow field (a two-channel map) which is used to warp the input image $x_a$ by means of a bilinear sampler $BS$. Here, the sampling procedure samples the pixels of $x_a$ at pixel coordinates determined by the flow field $f_{a\to b}$:

$$\tilde{x}_b(i, j, c) = x_a\{i + f_{a\to b}(i, j, 1), j + f_{a\to b}(i, j, 2), c\}, \tag{2}$$

where $\tilde{x}_b$ is the warped result representing the coarse output, $c$ denotes the channel of the image, and the curly brackets represent the bilinear interpolation which skips those positions with illegal values in the warping process. We use the $L2$ distance between the output $\tilde{x}_b$ and the ground truth $x_b$ as the objective function which is defined as follows:

$$L_{recon} = \mathbb{E}[\|\tilde{x}_b - x_b\|_2] \tag{3}$$

**NPG with Gazemap.** As shown in Fig. 2, we use the NPG output as an additional condition of the generation process. Jointly with the numerical gaze angle representation $r_a$, the pictorial gazemap $S$ is concatenated in a multimodal term to provide additional spatial and semantic information about the angle direction.

First, and differently from previous work in gaze redirection [6, 8], we compute the angle vector difference $\triangle r = r_b - r_a$, which is used as input instead of the absolute target angle to better preserve identity, similarly to [23]. Next, we generate the corresponding gazemap $S$ of the angles $r_a$ and $r_b$ by means of a synthesis process $F_s$ (details can be found below). Then, we concatenate $S_a$, $S_b$ and $\triangle r$ into a single term to get $\triangle r_s$:

$$\triangle r_s = [\triangle r, S_a, S_b]. \tag{4}$$

We detail below how we generate the gazemap ($F_s$). As shown in [16], our gazemap is also a two-channel Boolean image: one channel is for the eyeball which is assumed to be a perfect sphere, and the other channel is for the iris which is assumed to be a perfect circle. For an output map of size $m \times n$, with the projected eyeball diameter $2k = 1.2n$, the coordinates $(\mu, \nu)$ of the iris center can be computed as follows:

$$\begin{aligned} \mu &= \frac{m}{2} - k\cos\left(\arcsin\frac{1}{2}\right)\sin\phi\cos\theta \\ \nu &= \frac{n}{2} - k\cos\left(\arcsin\frac{1}{2}\right)\sin\theta, \end{aligned} \tag{5}$$

where the input gaze angle is $r = (\theta, \phi)$. The iris is drawn as an ellipse with the major-axis diameter of $k$, and the minor-axis diameter of $r\,|\cos\theta\cos\phi|$. Note that the synthesized gazemap only represents the gaze angle, without identity details of the specific eye sample. More visual examples of gazemaps can be found in the Supplementary Material.

### 3.2. Multi-task cGAN for Fine-grained Gaze Redirection

The warped result is inevitably blurry when using only the $L_2$ loss. Additionally, it also suffers from unwanted artifacts and unnatural distortions in the shape of the iris for large redirection angles. To remove these problems, we employ a generator $G$ to refine the output of the decoder. Instead of manipulating the whole image directly, we use $G$ to learn the corresponding residual image $R$, defined as the difference between the coarse output and the ground-truth. In this way, the manipulation can be operated with modest pixel modifications which provide high-frequency details, while preserving the identity information of the eye shape. The learned residual image is added to the coarse output of $Dec$:

$$\hat{x}_b = R + \tilde{x}_b. \tag{6}$$

where $\hat{x}_b$ represents the refined output.

**Conditional Residual Learning.** Learning the corresponding residual image $R$ is not a simple task as it requires the generator to be able to recognize subtle differences. Additionally, previous works [35, 6] indicate that introducing a suitable conditional information improves the performance of $G$. For this reason, we employ the input image $x_a$ and the head pose $h$ as conditional inputs for $G$. We also take the NPG output $\triangle r_s$ as input to provide stronger conditional information. The conditional residual image learning phase can be written as:

$$R = G(\tilde{x}_b, x_a, h, \triangle r_s). \tag{7}$$

Similarly to the coarse process, the image reconstruction loss, based on the $L2$ distance, is defined as follows:

$$L_{g\_recon} = \mathbb{E}\left[\|\hat{x}_b - x_b\|_2\right]. \tag{8}$$

The $L2$ loss penalizes pixel-wise discrepancies but it usually causes blurry results. To overcome this issue, we adopt the perceptual loss proposed in [10]. We use a VGG-16 network [19], pre-trained on ImageNet [4], which we denote as $\Phi$. The perceptual loss is defined as follows:

$$\begin{aligned} L_{g\_per} = \quad & \mathbb{E}\left[\frac{1}{h_j w_j c_j}\|\Phi_j(\hat{x}_b) - \Phi_j(x_b)\|_2\right] \\ + \quad & \mathbb{E}\left[\sum_{j=1}^{J}\|\Psi_j(\hat{x}_b) - \Psi_j(x_b)\|_2\right], \end{aligned}$$

where $\Phi_j(\cdot) \in \mathbb{R}^{h_j \times w_j \times c_j}$ is the output of the $j$-th layer of $\phi$. In our experiments, we use the activation of the 5th layer. $\Psi_j$ denotes the Gram matrix (for more details we refer the reader to [7]).

**Multi-task Discriminator Learning.** We use a multi-task discriminator in our model. Different from $G$, which is conditioned using multiple terms, the discriminator $D$ does not use them as input. Moreover, $D$ not only performs adversarial learning ($D_{adv}$) but also regresses the gaze angle ($D_{gaze}$). Note that $D_{adv}$ and $D_{gaze}$ share most of the layers with the exception of the last two layers. The regression loss is defined as follows:

$$\begin{aligned} L_{d\_gaze} &= \mathbb{E}\left[\|D_{gaze}(x_b) - r_b\|_2\right] \\ L_{g\_gaze} &= \mathbb{E}\left[\|D_{gaze}(\hat{x}_b) - r_b\|_2\right]. \end{aligned} \tag{9}$$

The adversarial loss for $D$ and $G$ is defined as:

$$\begin{aligned} \min_G \max_D L_{adv} = \mathbb{E}\left[log D_{adv}(x_b)\right] \\ + \mathbb{E}\left[log(1 - D_{adv}(\hat{x}_b))\right]. \end{aligned} \tag{10}$$

**Overall Objective Functions.** As aforementioned, we use $L_{recon}$ to train the encoder-decoder $Enc$ and $Dec$ to get

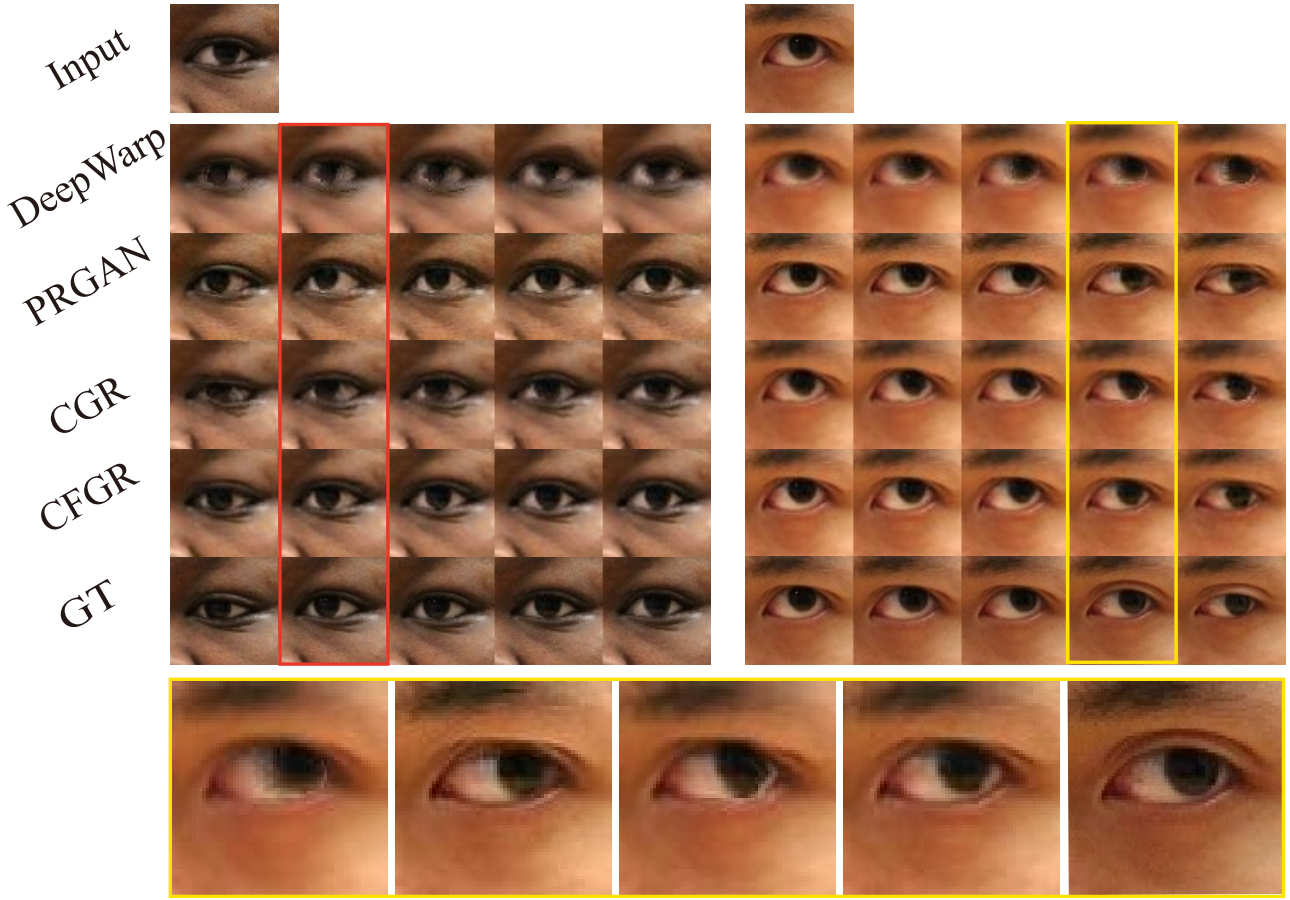


Figure 3. A qualitative comparison of different methods using redirection results with 10 different target angles ($\pm$ 15° head pose). The last row shows a magnification of the details marked with a yellow box in the previous rows, which correspond, from left to right, to DeepWarp, PRGAN, CGR, CFGR and GT.

the coarse-grained results. The overall objective function for $D$ is:

$$L_D = \lambda_1 L_{d\_gaze} - L_{adv}. \quad (11)$$

The overall objective function for $G$ is:

$$L_G = \lambda_2 L_{g\_recon} + \lambda_3 L_{g\_per} + \lambda_4 L_{g\_gaze} + L_{adv}. \quad (12)$$

$\lambda_1$ $\lambda_2$, $\lambda_3$ and $\lambda_4$ are hyper-parameters controlling the contributions of each loss term. Note that $L_G$ is used only to optimize $G$, but not to update $Enc$ and $Dec$.

## 4. Experiments

We first introduce the dataset used for our evaluation, the training details, the baseline models and the adopted metrics. We then compare the proposed model with two baselines using both a qualitative and a quantitative analysis. Next, we present an ablation study to demonstrate the effect of each component in our model, e.g., flow learning, residual image learning and the NPG module. Finally, we investigate the efficiency of our model. We refer to the full model as CFGR, and to the encoder-decoder with the only coarse-grained branch as CGR.

### 4.1. Experimental Settings

**Dataset.** We use the Columbia gaze dataset [20], containing 5,880 images of 56 persons with varying gaze directions and head poses. For each subject, there are 5 head directions ($[-30°, -15°, 0°, 15°, 30°]$) and 21 gaze directions ($[-15°, -10°, -5°, 0°, 5°, 10°, 15°]$ for the yaw angle and $[-10°, 0°, 10°]$ for the pitch angle, respectively). In our experiments, we use the same dataset settings of PRGAN [8]. In details, we use a subset of 50 persons (1-50) for training and the rest (51-56) for testing. To extract the eye region from the face image, we employ an external face alignment library (dlib [12]). Fixed $64 \times 64$ image patches are cropped as the input images for both training and testing. Both the RGB pixel values and the gaze directions are normalized in the range $[-1.0, 1.0]$. Other publicly available gaze datasets, e.g., MPIIGaze [33] or EYEDIAP [5], provide only low-resolution images and have not been considered in this evaluation.

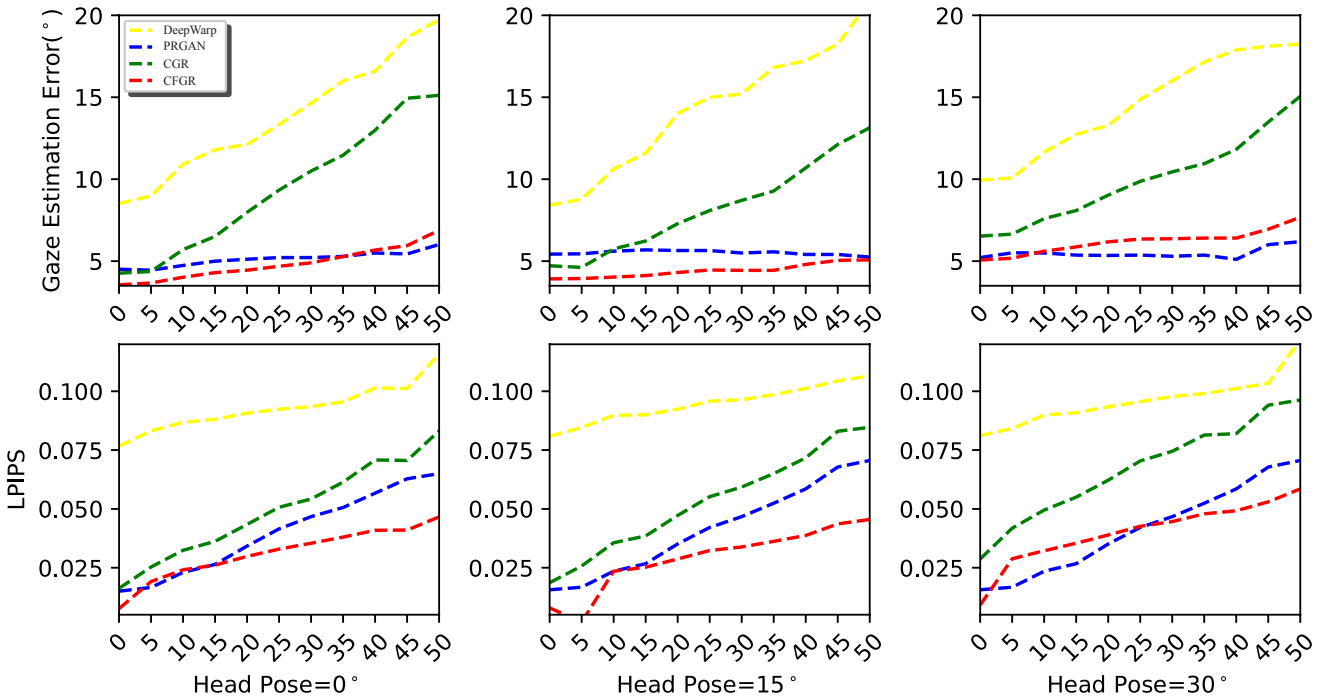


Figure 4. A quantitative evaluation of the gaze redirection results using three classes of head pose. First row: gaze estimation error. Second row: LPIPS scores. Lower is better for both metrics. Note that we combine the results of $\pm 15^\circ$ and $\pm 30^\circ$ head poses into $15^\circ$ and $30^\circ$.

**Training Details.** CGR is trained independently of the generator and the discriminator and it is optimized firstly, followed by $D$ and $G$. We use the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batch size is 8 in all the experiments. The learning rate for CGR is 0.0001. The learning rate for $G$ and $D$ is 0.0002 in the first 20,000 iterations, and then it is linearly decayed to 0 in the remaining iterations. $\lambda_1 = 5$, $\lambda_2 = 0.1$, $\lambda_3 = 100$ and $\lambda_4 = 10$ in our all experiments. The details of our network architectures can be found in the Supplementary Material.

**Baseline Models.** We adopt DeepWarp [6] and PRGAN [8] as the baseline models in our comparison. We use the official code of PRGAN* and train it using the default parameters. We reimplemented DeepWarp, as its code is not available. In details, different from the original DeepWarp, which is used only for a gaze redirection task with a single direction, we trained DeepWarp for gaze redirection tasks in arbitrary directions. Moreover, DeepWarp uses 7 eye landmarks as input, including the pupil center. However, detecting the pupil center is very challenging. Thus, we computed the geometric center among the 6 points as a rough estimation of the pupil center.

**Metrics.** How to effectively evaluate the appearance consistency and the redirection precision of the generated images is still an open problem. Traditional metrics, e.g., PSNR and MS-SSIM, are not correlated with the perceptual image quality [32]. For this reason, and similarly to PRGAN, we adopted the LPIPS metric [32] to compute the perceptual similarity in the feature space and evaluate the quality of redirection results. Moreover, we use GazeNet [33] as our gaze estimator and we pre-trained GazeNet on the MPIIGaze dataset to improve its gaze estimation.

*https://github.com/HzDmS/gaze_redirection

Table 1. Gaze-redirection quantitative evaluation. The scores represent the average of three head poses over ten redirection angles.

| Metric | LPIPS ↓ | Gaze Error ↓ |
|---|---|---|
| DeepWarp | 0.0946 | 14.18 |
| PRGAN | 0.0409 | 5.37 |
| CGR | 0.0565 | 9.19 |
| CFGR | **0.0333** | **5.15** |

## 4.2. Results

We first introduce the details of the qualitative and the quantitative evaluation protocols. For each head pose, we divide all the redirection angles into ten target groups by means of the sum of the direction differences in both pitch and yaw: $0^\circ$, $10^\circ$, $15^\circ$, $20^\circ$, $25^\circ$, $30^\circ$, $35^\circ$, $40^\circ$, $45^\circ$, $50^\circ$ (e.g., $0^\circ$ indicates that the angle differences between the target gaze and the input gaze are 0 in both the vertical and the horizontal direction). The test results of every group is used for the quantitative evaluation. Moreover, we select 10 redirection angles as the target angles for the qualitative evaluation: $[0^\circ, -15^\circ]$, $[10^\circ, -15^\circ]$, $[10^\circ, -10^\circ]$, $[10^\circ, -5^\circ]$, $[10^\circ, 0^\circ]$, $[10^\circ, 5^\circ]$, $[10^\circ, 10^\circ]$, $[10^\circ, 15^\circ]$, $[0^\circ,$

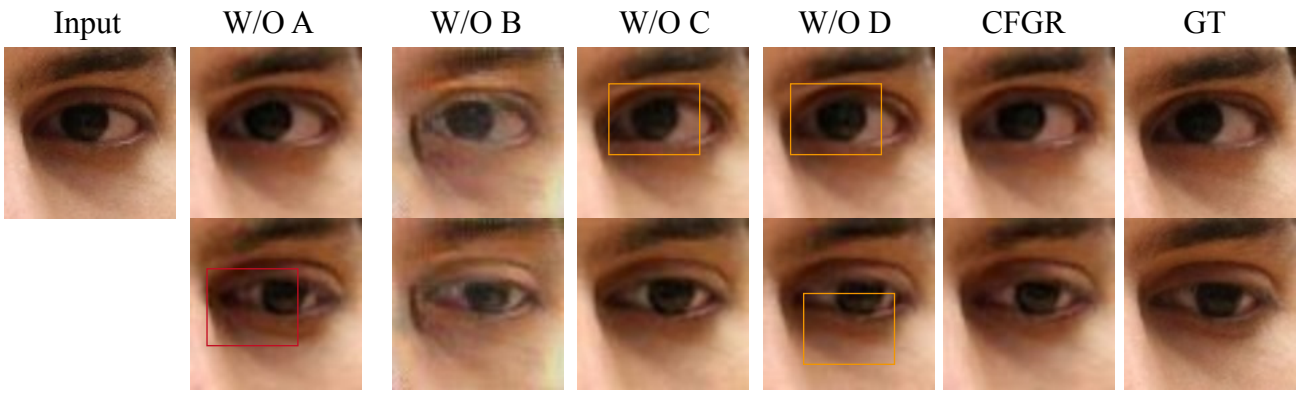


Figure 5. A qualitative comparison used in the ablation study. Red boxes: artifacts. Yellow boxes: unnatural shapes.

$15°]$, $[-10°, 15°]$.

**Qualitative Evaluation.** In the 5th row of Fig. 3, we show the redirection results of CFGR. The visually plausible results with respect to both the texture and the shape, and the high redirection precision, validate the effectiveness of the proposed model. Moreover, compared to CGR (without the refined generator module), we note that our refined model provides more detailed texture information and it eliminates unwanted artifacts and unnatural distortions of the iris shape.

As shown in the 2nd and in the 4th rows of Fig. 3, we observe that both DeepWarp and CGR redirect the input gaze with respect to the target angles, which demonstrates the ability of flow field in representing the correct spatial transformation. However, DeepWarp has several obvious disadvantages (marked with the yellow box in Fig. 3 and the corresponding zoom-in shown in the last row). For example, the generated textures are more blurry. In contrast, our coarse-grained CGR performs better. We attribute this to the fact that our encoder-decoder architecture with a bottleneck layer is better suitable for this task with respect to the scale-preserving fully-convolutional architecture adopted in DeepWarp.

As shown in the 3rd and in the 5th row of Fig. 3, both PRGAN and CFGR achieve high-quality redirection results with visual plausible textures and natural shape transformations for the iris. However, compared with CFGR, PRGAN suffers from two critical problems: (1) Lower image quality with a poor identity preservation (marked with a red box on the left); (2) Incorrect redirection angles and blurry boundaries causing distortion of the eyeball (marked with the yellow box and shown in the last row).

**Quantitative Evaluation.** In Fig. 4, we plot the gaze estimation errors and the LPIPS scores of different models. The three columns show the redirection results with respect to $0°$, $15°$ and $30°$ head pose angle, respectively. Note that we combine the results of $\pm 15°$ and $\pm 30°$ head poses into $15°$ and $30°$. It can be observed from the 1st row of Fig. 4 that CFGR achieves much lower gaze estimation error than DeepWarp and it is superior to PRGAN in most cases. Moreover, without the refined process, CGR has a much higher gaze error, especially for large gaze differences (e.g., $50°$).

Table 2. Results of the user study using three different head poses (with ten generated samples per pose). Every column sums to 100%. The rightmost column shows the overall performance.

| Head Pose | 0° ↑ | 15° ↑ | 30° ↑ | Average ↑ |
|---|---|---|---|---|
| DeepWarp | 7.32% | 10.18% | 5.69% | 7.73 % |
| PRGAN | 30.12% | 42.56 % | 45.79% | 39.49 % |
| CFGR | **62.56%** | **47.26%** | **48.52%** | **52.78%** |

The 2nd row of Fig. 4 shows the LPIPS scores. Here, we see that CFGR leads to much smaller scores than DeepWarp. Additionally, our model also has lower LPIPS scores than PRGAN, indicating that our method can generate a new eye image which is more perceptually similarly to the ground truth. However, CFGR has a higher gaze error or larger LPIPS scores in some cases, especially for redirection results with $30°$ head pose. Overall, as shown in Table 1, our approach achieves 0.0333 LPIPS score, lower than the 0.0946 of DeepWarp, the 0.0409 of PRGAN and it gets a 5.15 gaze error, lower than the 14.18 of DeepWarp and the 5.37 of PRGAN.

**User Study.** We conducted a user study to evaluate the proposed model with respect to the human perception. In details, we divided the gaze redirection results on the test data into three groups with respect to the head pose of the input image and we randomly selected 20 samples generated by each method for each group. Then, for each image, 10 users were asked to indicate the gaze image that looks more similar with the ground truth. Table 2 shows the results of this user study. We observe that our method outperforms PRGAN and DeepWarp in groups with $0°$, $15°$ and $30°$ head poses. Moreover, CFGR is selected as the best model on average, as shown in the final column of Table 2.

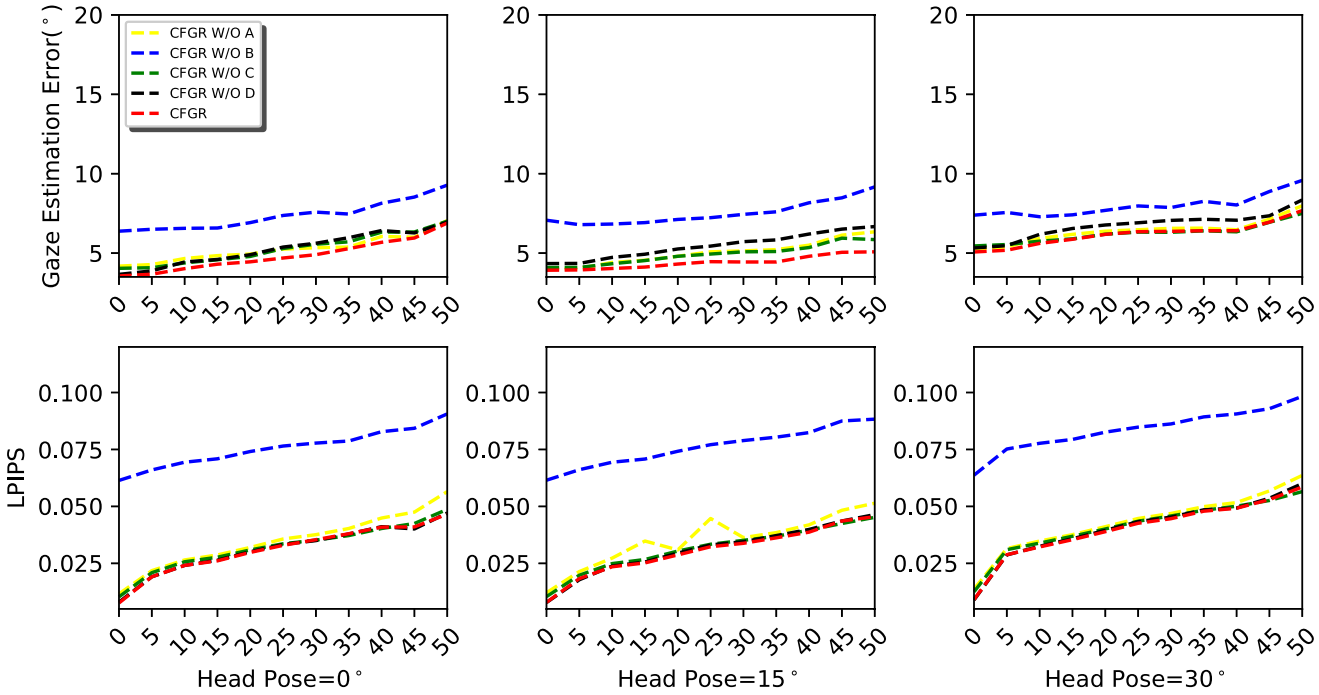


Figure 6. A quantitative evaluation used in the ablation study. The 1st row shows the gaze estimation error and the 2nd row the LPIPS scores. Our model ablated of the perceptual loss is called $A$, without the residual image learning is called $B$, without the flow field learning $C$, and without the pictorial gazemap guidance $D$.

### 4.3. Ablation Study

In this section we present an ablation study of the main components of our method. We refer to the full model without the perceptual loss as $A$. When we remove the flow learning in the encoder-decoder, this is called $B$. Removing the residual learning in the generator leads to model $C$, while removing the gazemap pictorial guidance gets $D$ (more details below).

**Perceptual Loss.** Fig. 5 shows that CFGR without the perceptual loss can generate results very close to the full model. However, some of these results have more artifacts (marked with a red box in the 2th column). Moreover, as shown in Fig. 6, the gaze estimation error and the LPIPS score are larger when removing this loss. Overall, the perceptual loss is helpful to slightly improve the visual quality and the redirection precision of the generated samples.

**Residual Learning.** We eliminate the residual term $R$ in Eq. 6 to evaluate its contribution. As shown in Fig. 5, the results are very blurry with a lot of artifacts. The quantitative evaluations in Fig. 6 are consistent with the qualitative results.

**Flow Learning.** Our encoder-decoder network predicts the flow field which is used to warp the input image for quickly learning the spatial shape transformation. As shown in Fig. 5, our full model achieves more natural results for the iris shape. Moreover, the quantitative results in Fig. 6 demonstrate the effectiveness of flow learning in improving the redirection precision.

**Gazemap in NPG.** When removing the gazemap (see the 6th column in Fig. 5), the visual results present more shape distortions compared with the full model. Moreover, the quantitative results in Fig. 6 demonstrate the effect of the gazemap in improving the redirection precision.

## 5. Conclusion

In this paper we presented a novel gaze redirection approach based on a coarse-to-fine learning. Specifically, the encoder-decoder learns to warp the input image using the flow field for a coarse-grained gaze redirection. Then, the generator refines this coarse output by removing unwanted artifacts in the texture and possible distortions of the shape. Moreover, we proposed an NPG module which integrates a pictorial gazemap representation with the numerical angles to further improve the redirection precision. The qualitative and the quantitative evaluations validate the effectiveness of the proposed method and show that it outperforms the baselines with respect to both the visual quality and the redirection precision. In future work we plan to extend this approach to the gaze redirection task in the wild.

**Acknowledgement:** This work is sponsored by Shanghai Pujiang Program (No.19PJ1402000), in part by Science and Technology Commission of Shanghai Municipality Project (19511120700) and Shanghai Engineering Research Center of AI & Robotics, China and Engineering Research Center of AI Robotics, Ministry of Education, China.

## References


[1] Michael Banf and Volker Blanz. Example-based rendering of eye movements. *Computer Graphics Forum*, 28(2):659–666, 2009.

[2] Juntong Cheng, Yi-Ping Phoebe Chen, Minjun Li, and Yu-Gang Jiang. TC-GAN: Triangle cycle-consistent gans for face frontalization with facial features preserved. In *ACM Multimedia*, 2019.

[3] 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 *CVPR*, 2018.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Symposium on Eye Tracking Research and Applications*, 2014.

[6] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *ECCV*, 2016.

[7] Leon Gatys, Alexander Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

[8] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *ICCV*, 2019.

[9] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. AttGAN: Facial attribute editing by only changing what you want. *IEEE Trans. on Image Processing*, 28(11):5464–5478, 2019.

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[11] Harsimran Kaur and Roberto Manduchi. EyeGAN: Gaze-preserving, mask-mediated eye image synthesis. In *WACV*, 2020.

[12] Davis King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(7), 2009.

[13] Daniil Kononenko and Victor Lempitsky. Learning to look up: Realtime monocular gaze correction using machine learning. In *CVPR*, 2015.

[14] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. *arXiv preprint arXiv:1904.09709*, 2019.

[15] Yahui Liu, Marco De Nadai, J. Yao, N. Sebe, Bruno Lepri, and Xavier Alameda-Pineda. Gmm-unit: Unsupervised multi-domain and multi-modal image-to-image translation via attribute gaussian mixture modeling. *ArXiv*, abs/2003.06788, 2020.

[16] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *ECCV*, 2018.

[17] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[18] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. GANimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, 128:698–713, 2020.

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

[20] Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. Gaze locking: Passive eye contact detection for human-object interaction. In *ACM Symposium on User Interface Software and Technology*, 2013.

[21] H. Tang, X. Chen, W. Wang, D. Xu, J. J. Corso, N. Sebe, and Y. Yan. Attribute-guided sketch generation. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–7, 2019.

[22] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Gazedirector: Fully articulated eye gaze redirection in video. *Computer Graphics Forum*, 37(2):217–225, 2018.

[23] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5914–5922, 2019.

[24] Ruizheng Wu, Xin Tao, Xiaodong Gu, Xiaoyong Shen, and Jiaya Jia. Attribute-driven spontaneous motion in unpaired image translation. In *ICCV*, 2019.

[25] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. *arXiv preprint arXiv:2003.05905*, 2020.

[26] Weidong Yin, Ziwei Liu, and Change Loy Chen. Instance level facial attributes transfer with geometry-aware flow. In *AAAI*, 2019.

[27] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *CVPR*, 2019.

[28] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. *arXiv preprint arXiv:1911.06939*, 2019.

[29] Jichao Zhang, Jingjing Chen, Hao Tang, Wei Wang, Yan Yan, Enver Sangineto, and Nicu Sebe. Dual in-painting model for unsupervised gaze correction and animation in the wild. In *ACM Multimedia*, 2020.

[30] Jichao Zhang, Yezhi Shu, Songhua Xu, Gongze Cao, Fan Zhong, Meng Liu, and Xueying Qin. Sparsely grouped multi-task generative adversarial networks for facial attribute manipulation. In *ACM Multimedia*, 2018.

[31] Jichao Zhang, Fan Zhong, Gongze Cao, and Xueying Qin. ST-GAN: Unsupervised facial image semantic transformation using generative adversarial networks. In *ACML*, pages 248–263, 2017.

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[33] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2017.

[34] Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li, and Ran He. A survey of deep facial attribute analysis. *IJCV*, pages 1–33, 2020.

[35] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.