Learning Gaze-aware Compositional GAN from Limited Annotations

NEREA ARANJUELO, Fundación Vicomtech, Basque Research and Technology Alliance , Spain

SIYU HUANG, Clemson University, USA

IGNACIO ARGANDA-CARRERAS, University of the Basque Country, Ikerbasque, Donostia International Physics Center, and Biofisika Institute, Spain

LUIS UNZUETA, Fundación Vicomtech, Basque Research and Technology Alliance , Spain

OIHANA OTAEGUI, Fundación Vicomtech, Basque Research and Technology Alliance, Spain

HANSPETER PFISTER, Harvard John A. Paulson School of Engineering and Applied Sciences, USA
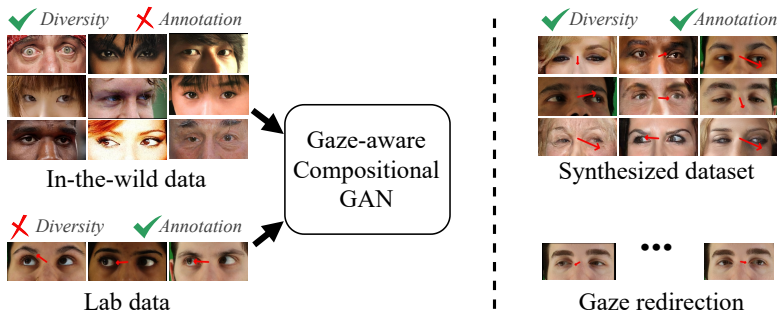
DONGLAI WEI, Boston College, USA

Fig. 1. Left: The proposed Gaze-aware Compositional GAN model can leverage the limited gaze annotation in the lab data and diverse appearance in the unlabeled data. Right: Applications include gaze-annotated image dataset generation and gaze redirection.

Gaze-annotated facial data is crucial for training deep neural networks (DNNs) for gaze estimation. However, obtaining these data is labor-intensive and requires specialized equipment due to the challenge of accurately annotating the gaze direction of a subject. In this work, we present a generative framework to create annotated gaze data by leveraging the benefits of labeled and unlabeled data sources. We propose a Gaze-aware Compositional GAN that learns to generate annotated facial images from a limited labeled dataset. Then we transfer this model to an unlabeled data domain to take advantage of the diversity it provides. Experiments demonstrate our approach's effectiveness in generating within-domain image augmentations in the ETH-XGaze dataset and cross-domain augmentations in the CelebAMask-HQ dataset domain for gaze estimation DNN training. We also show additional applications of our work, which include facial image editing and gaze redirection.

CCS Concepts: • **Computing methodologies** → *Neural networks*; **Image processing**; • **Human-centered computing** → Interaction techniques.

Additional Key Words and Phrases: Gaze estimation, synthetic data, GAN, generative, DNN, domain transfer

1 INTRODUCTION

Gaze estimation is an essential task in computer vision for many applications, including but not limited to human-computer interaction, virtual reality, and the automotive industry. Recent works in gaze estimation have demonstrated that Deep Neural Networks (DNNs) are more robust and accurate than traditional methods [Cheng et al. 2021]. However, obtaining enough quality data is essential for training DNNs for gaze estimation. Despite the abundance of facial images through the Internet and open large-scale datasets [Lee et al. 2020], obtaining gaze-annotated images remains a major challenge. Manually annotating existing images is tedious and prone to errors, as accurately determining where a person is looking is difficult. Capturing gaze-annotated data is labor-intensive and typically requires specialized equipment and long capturing sessions, where a volunteer is asked to direct their gaze to various points [Zhang et al. 2020].

Given the high value of each captured sample, it is crucial to explore alternative methods for obtaining them. One potential solution is to use synthetic data [Nikolenko 2021]. Data might be generated using 3D simulated environments. However, this approach involves its own difficulties, such as the need for additional processes to bridge the gap between the simulated and real data domains. Synthetic data generation and augmentation using generative models might be a promising alternative. Generative adversarial networks (GAN) can produce high-quality images with a high level of realism [Karras et al. 2019] with the added benefit of controlling the generated data by editing latent input vectors or conditioning the model on some input information [Shu et al. 2021].

In this paper, we present a Gaze-aware Compositional GAN learned from limited annotations. As shown in Fig. 1, the proposed model can be used for both discriminative tasks like gaze estimation by synthesizing gaze-annotated facial images and generative tasks like gaze redirection. Our approach relies on a composition-based gaze-aware generative model to augment a limited annotated dataset. Our method consists of two stages. First, we train a compositional GAN to generate gaze-aware images using a small annotated training dataset. Second, we transfer the model to a bigger in-the-wild unlabeled dataset. Then, we can generate within- and cross-domain new images to augment the limited annotated data. Our code is available at https://github.com/naranjuelo/GC_GAN.

The main contributions of this paper are:

- We introduce a novel approach for acquiring annotated gaze estimation data from limited annotations, leveraging the strengths of various labeled and unlabeled data sources.
- We develop a Gaze-aware Compositional GAN that generates realistic synthetic images with a user-specified gaze direction using a limited annotated dataset. We transfer the gaze-aware generative model to an unlabeled dataset to exploit the variance of in-the-wild captured data.
- We validate our method by demonstrating its ability to generate previously unseen within-domain and cross-domain image augmentations that boost the accuracy of a gaze estimation DNN. We show our method can be used for different applications, such as data generation, image editing, data augmentation, and gaze redirection.

2 RELATED WORK

2.1 Gaze Estimation from Facial Images

Traditional model-based methods for estimating gaze rely on a 3D eye model and use detected features such as the pupil center or the corneal reflections to estimate the gaze [Ishikawa 2004]. Recent

appearance-based approaches use DNNs to estimate the gaze using images as input [Abdelrahman et al. 2022; Cheng and Lu 2022; Zhang et al. 2020]. These approaches adapt better to challenging in-the-wild scenarios but strongly rely on large amounts of varied data for training. Public available datasets might alleviate this need. Some of these datasets are captured in controlled environments (e.g., a laboratory) and typically include relatively few subjects but a wide range of balanced gaze samples [Xia et al. 2020; Zhang et al. 2020]. Datasets obtained in more natural conditions typically include a wider variety of subjects but lack the controlled wide range of gazes [Kellnhofer et al. 2019; Zhang et al. 2015]. Due to the variability in person-specific appearance and the influence of the setup (e.g., face perspective), gaze estimation DNNs are frequently fine-tuned before deployment with some samples captured in the target setup [Arar et al. 2016; Masko 2017; Park et al. 2019]. Typically, this involves a tedious process where users sequentially look at different targets. Thus, DNNs are often retrained with a limited number of images. Due to the challenges of getting accurately gaze-annotated data, synthetic data emerge as an alternative to data capturing.

2.2 Gaze-aware Facial Image Generation

Realistic eye image synthesis has been extensively studied for its importance in various applications. There are two main approaches: 3D eye modeling and image generation using generative models.

Graphics-based methods produce high-resolution images based on a 3D model of the human eye region [Wood et al. 2016]. Although gaze directions and head poses can be controlled in these approaches, using artificial eye texture can make the generated images appear unrealistic and widen the discrepancies between generated and target data, leading to lower accuracy when training DNNs with these unrealistic data. Some works propose to reduce the domain discrepancies by leveraging GANs to enhance the realism of synthetic samples before using them for DNN training [Shrivastava et al. 2017]. GazeGAN [Sela et al. 2017] employs unpaired image-to-image translation to transform images between domains. These methods involve generating synthetic data from 3D models, which can be time-consuming due to 3D asset design and rendering.

Some recent works use Neural Radiance Fields (NeRFs) [Mildenhall et al. 2021] to synthesize photoreal images with gaze control. In [Li et al. 2022], mesh-based and volumetric reconstruction are combined to synthesize the eye region. Still, it lacks modeling of facial expression-induced eye deformations, requires a complex image capture setup, and struggles with details like eyelashes.

To overcome the aforementioned challenges, our work generates realistic images from latent vectors using GANs. The state-of-the-art GANs [Brock et al. 2018; Karras et al. 2021] can generate realistic images without a 3D simulation environment. Our work is inspired by the recent composition-based generation SemanticStyleGAN [Shi et al. 2022]. Although many of these GANs focus on facial image generation, gaze-aware data generation remains an unexplored challenge.

2.3 Dataset Generation for Gaze Estimation

Given the limited and hard-to-obtain gaze data, GANs might be convenient to augment the existing data. Many works have attempted to use GANs to achieve controllable image generation or editing. This can be tackled by learning a model to manipulate the latent space of a pre-trained GAN [Härkönen et al. 2020; Shen et al. 2020] or training with additional supervision to learn a more disentangled latent space [Deng et al. 2020; Shi et al. 2022]. These works are often applied to face attributes edition, such as hair or skin, but not gaze direction. Furthermore, most of the generated images tend to look forward as it is the most common gaze in typical training datasets [Karras et al. 2019; Lee et al. 2020]. GANs also have limitations, based on what they have seen during training [Jahanian et al. 2019], so the training data domain influences their generation capabilities.

Specific studies focus on gaze correction or redirection. InterpGaze [Xia et al. 2020] generates intermediate gaze images given two samples using an encoder, a controller, and a decoder. [Chen
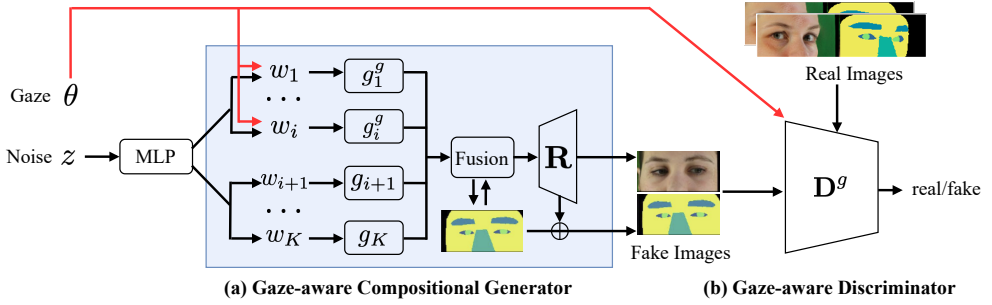
(a) Gaze-aware Compositional Generator   **(b) Gaze-aware Discriminator**

Fig. 2. The proposed Gaze-aware Compositional GAN model (GC-GAN). Building upon the SemanticStyle-GAN [Shi et al. 2022], we first group facial components into gaze-related $\{w_1, \ldots, w_i\}$ and gaze-unrelated $\{w_{i+1}, \ldots, w_K\}$. We extend (a) the Local Generators ($\{g_j^g\}_1^i$) in the generator and the discriminator D^g to condition on the input gaze θ.

et al. 2021] also uses an encoder-decoder architecture, followed by a refinement generator. These works are limited to redirecting the gaze of a cropped eye region. Some works also learn to blend this crop into the face, such as DeepWarp [Ganin et al. 2016] and GazeGAN [Zhang et al. 2019], but are constrained to gaze correction or limited movements. EyeGAN [Kaur and Manduchi 2020] uses a segmentation mask instead of an RGB image to generate an image for a gaze direction. However, it is limited to tight grayscale eye-crop images. Our approach is not limited to gaze redirection but allows gaze-aware image generation. Using a composition-based architecture, we can control specific face components for data augmentation. Finally, combining datasets from different domains and with different strengths enhances the variability of images our method generates.

3 METHOD

In a typical GAN, a generator model is trained to map a noise vector z to an image x. Our proposed method learns to generate an eyes' region image x and a corresponding segmentation mask y given a vector z and a target gaze direction θ the eyes should follow, as defined below:

$$G : (z, \theta) \longrightarrow (x, y), \tag{1}$$

where $y \in \{0, 1\}^{HxWxK}$, being H and W the image dimensions, and K the number of segmented face components. Our method allows training on labeled and unlabeled domains, which adds the challenge of having no annotations to supervise the synthesized images' gaze in training.

3.1 Gaze-aware Compositional GAN

We have the following three observations: (1) Some facial components are related to the gaze (*e.g.*, iris), while others are not (*e.g.*, nose). (2) Our facial image generator needs to be compositional, as we need the gaze-related facial components to remain the same during the gaze-invariant data augmentation for the gaze estimation task and to be changed alone during the gaze redirection editing task. (3) Our discriminator needs to penalize the sample if it doesn't match the conditioned gaze direction in addition to being not realistic. Thus, we designed the following model components.

Overall framework. To ensure the composability of our GAN, we base our model on the state-of-the-art SemanticStyleGAN model [Shi et al. 2022]. We extend SemanticStyleGAN to support effective gaze conditioning while generating realistic images and learning from labeled and unlabeled domains. As illustrated in Fig. 2, our Gaze-aware Compositional GAN model (GC-GAN) is divided into three main components: noise to latent vectors mapping, Gaze-aware Compositional Generator for image generation, and Gaze-aware Discriminator for discriminating images during training. The

(a) Gaze-aware Local Generator (GLG)   **(b) Discriminator (first training stage)**
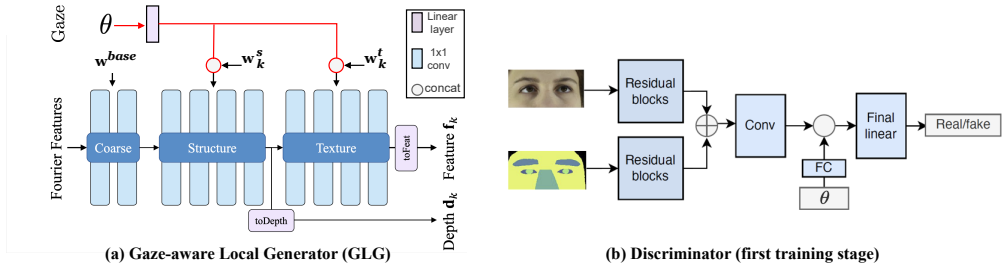
Fig. 3. Architectures of the GC-GAN. Gaze-aware Local Generator (GLG) (a): the red lines are additions compared to [Shi et al. 2022]. The discriminator in the first training stage (b): outputs from residual blocks are summed, while features from gaze direction concatenated.

input vector z is mapped to an intermediate latent code $w \sim W$ using an 8-layer MLP to better model the non-linearity of the data distribution, similar to [Karras et al. 2019]. To obtain a disentangled latent space for different face components, the latent code w is divided into K local latent codes and an additional base latent code $\mathbf{w}^{base}$, common for all the face components. These latent codes are processed by the Gaze-aware Compositional Generator. This module has K generators, each responsible for generating feature maps for a specific face component given a local latent code w_k. All the output feature maps are fused and fed to the final generator, which generates synthetic images and the corresponding segmentation masks. During training, the generated images and masks are fed to the discriminator, as well as real samples from training data. Gaze vectors are also input to the discriminator when available. The main modules of the model are detailed as follows.

Gaze-aware generator and discriminator. For image generation, as the Render Net is gaze agnostic, we only need to condition the gaze direction for the Local Generators of gaze-related facial components, $\{g_j^g\}_1^i$. For gaze-related features, we condition the Local Generators with the gaze direction feature. We refer to these generators as Gaze-aware Local Generators (GLG). Fig. 3 (left) shows the GLG architecture for a specific facial component k. Each local generator is formed by modulated 1×1 convolution layers with latent code-conditioned weights, and input Fourier features f_p for position encoding. The input latent codes for each local generator are the base latent code $\mathbf{w}^{base}$, and the face component-specific w_k. The latent code w_k is divided into shape and texture latent vectors, w_k^s and w_k^t [Shi et al. 2022]. The target gaze direction θ is also input to the GLGs so that we can control the gaze of generated samples. The input gaze θ is defined as yaw and pitch angles (θ_y, θ_p), which are fed to a fully connected layer for mapping them to an adequate 64-dimensional space before fusion. The output of this layer is concatenated to the component-specific latent vectors, and the extended latent vectors are fed to a series of modulated 1×1 convolutions and final fully connected layers. The output of the GLG is a 1-channel pseudo-depth d_k and 512-channel feature map f_k. The hidden layers have 64 channels. The local generators unrelated to gaze follow the same architecture except for the gaze branch, which we discard.

Regarding the discriminator, our generated samples are discriminated based on the joint distribution of the image, the mask, and the gaze (Fig. 3, right). For the image and mask, the discriminator has two convolution branches whose outputs are summed up. Each branch is formed of residual blocks, similar to [Karras et al. 2019; Shi et al. 2022], which convert the input 256×256 images to 4×4 feature maps. The mini-batch standard deviation of the summed feature maps is concatenated, and 3×3 kernel convolution is applied before reshaping to a 1-dimensional vector and adding the data from the gaze. The gaze is included as a third input to the discriminator. A linear layer maps

the gaze to a 64-dimensional space before being concatenated with feature maps from the dual branch. These maps are processed by the final linear layer, classifying the input data as real or fake.
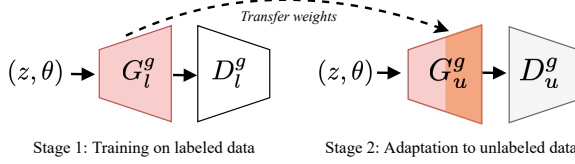

Fig. 4. Two-stage training: (1) training on limited annotated data, (2) adaptation to unlabeled data.

3.2 Two-stage Training

We first train the GC-GAN model on the labeled data and then transfer the facial appearance from the unlabeled data with the domain adaptation approach (Fig. 4).

Stage 1: Training on labeled data. During the first stage, the GC-GAN (Section 3.1) is trained by minimizing the loss

$$L_{s1} = \lambda_l L_l + \lambda_r L_r + \lambda_p L_p + \lambda_m L_m + \lambda_s L_s \,, \tag{2}$$

where L_l is non-saturating logistic loss [Creswell et al. 2018], L_r is R1 regularization loss [Mescheder et al. 2018], L_p is path length regularization loss [Karras et al. 2019], L_m and L_s are mask and segmentation regularization loss [Shi et al. 2022]. Each loss is weighted with the corresponding λ (Section 4). However, due to the limited amount of labeled data, it is hard for such a trained model to generalize well to new in the wild images that are challenging to have accurate gaze annotations.

Stage 2: Adaptation to unlabeled data. To adapt GANs to the target domain, it is a common practice to freeze the lower-level layers of the generator module [Benaim and Wolf 2018]. The challenge here is to find out which module to freeze and, additionally, how to handle the lack of annotations in the unlabeled domain. In our case, the target domain lacks the label information that is used in the source domain, which requires the modification of the discriminator model.

Freezing Gaze-aware Compositional Generator. To transfer the model to the appearance distribution from a new domain, we rely on the hypothesis that samples with the same segmentation mask share the same gaze direction. A specific latent vector w and input gaze direction θ, fed to the gaze-aware generators module, results in a fused coarse feature map and mask, which are then refined in the Render Net. Our goal is that the same latent vector generates images in both dataset domains that share the same gaze direction and coarse features (e.g., pose), but domain-specific appearances. For that purpose, we initialize the model with the pretrained weights from the first stage, freeze the GLG and the first block of the rendering generator, and fine-tune the rest of the model. Random input gazes are fed to the generator.
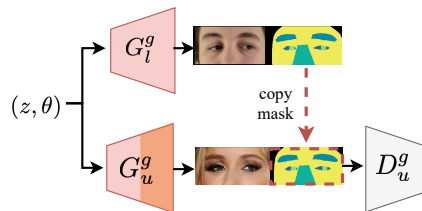

Fig. 5. Stage 2: Adaptation to unlabeled data. The generated image is combined with the mask generated by the pretrained stage-1 model given the same latent and gaze vectors.

Modifying Gaze-aware Discriminator. There are no available gaze annotations to use as input along with real images and masks, so we constrain the gaze requirement by using the segmentation mask. The discriminator has the same architecture as in Section 3.1 without the input gaze branch. The discriminator receives real pairs of RGB images and the corresponding segmentation masks, and the same with synthetic pairs. In the synthetic pairs, we use the synthetic image generated by the generator given a latent vector w and a target gaze direction θ, but an expected segmentation mask for that latent vector instead of the generated one. The expected mask is generated by the previously pretrained model, as shown in Fig. 5. The model trained in the first stage has already learned how to model the relation between the latent code and the gaze in the generated images. Consequently, using it to generate the masks for the discriminator's input pairs forces the learning GAN to generate images that fit the given masks and, hence, the input gaze direction.

We also add a mask loss L_g, which is minimized together with the same losses as the first stage. This loss helps to preserve the same generated mask for the same latent vector and, thus, the same gaze direction. The new mask loss measures the mean squared error between the generated mask with the updated weights and the mask generated using the pretrained weights given the same latent and gaze vector. The final loss of the second stage is defined as follows:

$$L_{s2} = L_{s1} + \lambda_g L_g, \tag{3}$$

where L_{s1} are the losses defined in Equation 2 and L_g is the mask loss for gaze preservation.

3.3 Applications

Gaze-aware facial image synthesis. The composition-based architecture promotes a disentangled and interpretable latent space for image editing. After model training, given an intermediate latent vector w and a target gaze θ we can generate a synthetic sample. We can redirect the gaze of this sample by modifying the input gaze direction θ. Regarding the semantic facial components, we can modify the specific k_{th} component by varying the corresponding local latent vector w_k^s or w_k^t. We use cubic spline interpolation to smoothly move from a local latent vector to a different one.

Data augmentation for gaze estimation. Fig. 6 shows different ways for generating data we propose: data generation in different domains using the same latent and gaze vector, redirecting the gaze, and modifying specific facial components.
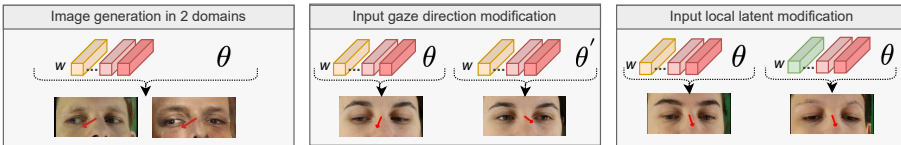

Fig. 6. Design choices for data augmentation with the trained GC-GAN model.

Gaze-aware facial image editing. To edit the gaze direction of a facial image, we need to map it to the learned latent space through image inversion. This process involves optimizing a latent vector to generate an image that closely resembles the desired target image. We use the loss:

$$L_{inv} = \lambda_p L_p + \lambda_d L_d + \lambda_m L_m, \tag{4}$$

where the perceptual loss L_p measures the high-level similarity between the generated and target images computing the L2 distance between intermediate image features using a pretrained VGG19 network [Zhang et al. 2018]. The L_d computes the pixel-wise MSE loss between both images. Finally, the L_m computes the difference between the evaluated latent vector and a mean latent vector to

prevent significant deviations. The mean latent vector is computed by averaging multiple latent vectors from random z vectors. All losses are weighted with the corresponding λ (Section 4).

4 EXPERIMENTS

We thoroughly examine our proposed framework through qualitative and quantitative experiments to showcase its effectiveness in generating data augmentations, as well as gaze-controlled images.

4.1 Datasets and Implementation Details

Datasets. We use the ETH-XGaze dataset [Zhang et al. 2020] as the primary source of gaze-labeled data. It contains images of 110 subjects captured from 18 cameras in a controlled environment. We randomly select a subset of 8 subjects for training (14,464 images, subject IDs: 0, 3, 28, 29, 52, 55, 81, 83) and 2 for testing (3,920 images, subject IDs: 19, 24) for the four most frontal cameras (camera IDs: 1, 2, 3, 8). While the ETH-XGaze dataset features a wide and balanced gaze range, the variety and naturalness of individuals captured in controlled environments may be limited. To address this limitation, we use the CelebAMask-HQ dataset [Lee et al. 2020] as the unlabeled dataset, comprising 30,000 in-the-wild celebrity images.

Fig. 7. Preprocessing of a sample from the CelebAMask-HQ dataset: face and face landmarks are detected for image normalization and eyes' region image crop and segmentation.

Data preprocessing. We preprocess the data from the ETH-XGaze and CelebAMask-HQ datasets. It is common to use image crops containing one or both eyes for gaze estimation DNNs [Masko 2017; Porta et al. 2019; Shrivastava et al. 2017; Sinha et al. 2021]. We use the eyes' region crop for our experiments. For models using the entire face, our approach can be extended to the face. We preprocess the images to get a normalized 256×256 eyes' region image crop and the corresponding segmentation mask. We first detect the faces in the images [Bazarevsky et al. 2019] and their face landmarks [Kartynnik et al. 2019]. We normalize each image by rotating it to eliminate any roll angle due to the head pose, centering it based on the face center, and scaling it to have 1.7 times the eyes' region width, considering the furthest eye corners horizontally. We define the face center as the average point considering the average of the eyes' coordinates and the mouth's coordinates. The eyes' region is the facial image's upper half. We use the face landmarks to generate the corresponding segmentation mask, which includes the following categories: background, face, iris, sclera, eyebrows, and nose. Fig. 7 shows an example for a CelebAMask-HQ dataset sample.

Training. We first train the composition-based GAN with the ETH-X Gaze subset and then transfer the model in the second stage to the CelebAMask-HQ dataset. We set K to 6 for the considered number of components in the image. We have 4 non-conditioned local generators (for background, face, eyebrows and nose) and 2 GLGs (for sclera and iris). Random input gaze angles are fed to the model during training, sampled from the real gaze range of the labeled data. We set a dimensionality of 512 for z and w, and the image size is 256×256. Similar to StyleGAN2, we use style mixing regularization [Karras et al. 2019] and leaky ReLU activations. We empirically set the first stage's λ hyperparameters to $\lambda_l = 1$, $\lambda_r = 10$, $\lambda_p = 2$, $\lambda_m = 100$, and $\lambda_s = 500$. In the second stage, λ_g is set to 100. We train our model for 250k iterations in the first stage and 470k iterations in the second. We use the ADAM optimizer [Kingma and Ba 2014] with $\beta_1 = 0$, $\beta_2 = 0.99$, and 8 as the minibatch. We implement our framework using PyTorch 1.7 and we use an Nvidia GPU GeForce RTX 3090.

4.2 Data Augmentation for Gaze Estimation

To assess the effectiveness of our method for data augmentation, we train a gaze estimation DNN with and without augmented data, respectively, and evaluate the DNN accuracy on the test set. We consider four configurations: (1) a baseline configuration with common augmentations with geometric and color modifications (e.g., CLAHE, shift and scale perturbations), (2) a variation of the baseline configuration that only employs color modifications (e.g., brightness variations), (3) GAN-based within-domain augmentations, and (4) GAN-based within and cross-domain augmentations.

Regarding the gaze estimation DNN training, we train an off-the-shelf ResNet-50 [He et al. 2016] as in [Zhang et al. 2020]. We train the DNN on 224×224 eyes' region images to estimate gaze direction as yaw and pitch angles. We use the ADAM optimizer with a starting learning rate of 1e-3, a batch size of 50 and train for 25 epochs with learning rate decay every ten epochs. We select the best training epoch for testing based on the validation set (15% of the training set). We also train two other state-of-the-art DNNs to see the suitability of our synthetic data with different architectures. Specifically, we use the training configurations and implementations provided for the L2CS [Abdelrahman et al. 2022] and the transformer-based GazeTR [Cheng and Lu 2022].

For GAN-generated augmented samples, we first embed the ETH-XGaze subset samples into the closest latent vectors. New samples are then generated through iterative modification of unconditioned face components and redirecting the gaze vector of some of the samples (Section 3.3). We generate 8,000 synthetic images for the within-domain augmentations and further augment the training set by an additional 5% with the cross-domain augmentations. We use the same inverted images' latent vectors to generate the synthetic images in the CelebAMask-HQ dataset domain.

To evaluate the accuracy of the gaze predictions, we compute the gaze angular error (°) as:

$$L_{gaze} = \frac{g \cdot g'}{||g|| \, ||g'||} \tag{5}$$

where $g \in \mathcal{R}^3$ is the ground-truth gaze and $g' \in \mathcal{R}^3$ the predicted gaze vector. The trained DNNs predict pitch and yaw angles, which are converted to 3D vectors (no roll angle [Zhang et al. 2015]).

Table 1 shows the error obtained by the DNN in the test set when trained with the different augmented sample configurations, as well as when trained without augmentations. While all augmentations lead to an error drop, the reduction is significantly higher when using GAN-based augmentations. Additionally, we observe that even though cross-domain augmentations belong to a different domain, they help to improve the model's generalization capability.

Table 1. Error (°) in the ETH-XGaze test set when trained with limited data and different kinds of augmentations and DNN architectures [Abdelrahman et al. 2022; Cheng and Lu 2022; Zhang et al. 2020].

Augmentation	Number of images			Error ↓		
	ETH-X (r)	ETH-X (s)	CELEB (s)	ResNet	L2CS	GazeTR
No Aug	14,464	0	0	4.54	4.07	5.11
Color Aug	14,464	0	0	4.47	3.67	4.89
Color+Geometry Aug	14,464	0	0	4.52	3.39	4.81
Ours (in-domain)	14,464	8,000	0	4.22	3.16	5.04
Ours (in- and cross-domain)	14,464	8,000	1,123	**3.86**	**3.13**	**4.65**

Fig. 8 shows qualitative results of the different augmentations, as evaluated in Table 1. The images on the left have undergone geometry and color modifications, as used in the baseline augmentations configuration. The second pair of images are within-domain augmentations for the same subject. The images on the right depict a pair of synthetic images in the ETH-XGaze and CelebAMask-HQ

Fig. 8.   Different augmentations evaluated for DNN training. Left to right: color and geometric augmentations, within-domain augmentations, and cross-domain augmentations.

data domains, generated using the same latent vector. It can be observed that despite differing appearances, images within each pair maintain consistent gaze direction.

4.2.1 Ablation Studies: data generation for gaze estimation DNN training

Data augmentation: in-domain synthesis. We evaluate how varying the number of GAN-generated augmentations affects the accuracy of the gaze estimation DNN. We repeat the experiment in Section 4.2 with the DNN from [Zhang et al. 2020], varying the number of synthetic samples in the training set. Fig. 9a shows the obtained mean error in the estimated gaze for the different training sets adding within-domain augmentations. When training with no synthetic samples, the DNN obtains an error of 4.53° in the test set. When we add some synthetic samples to the training set, the error is similar (4.52°), but as we increase the number of synthetic samples (more than 5,000 images), the error decreases to 3.90°. The error drop shows that the augmentations are helpful.

Data augmentation: cross-domain synthesis. We increase each training set a 5% by selecting the images with the highest confidence. We compute this confidence score based on the MSE between the generated masks in both dataset domains. We repeat the DNN training with different sets of augmented data. Fig. 9b shows the results in the test set when adding within-dataset and cross-dataset augmentations to the training set. It can be seen that adding cross-dataset augmentations helps in reducing the error. Adding images from a new domain increases the accuracy of the DNN, but adding an excessive amount of images does not seem to necessarily lead to further improvement (the minimum test error is 3.86°, and it is achieved with 9,000 synthetic images).

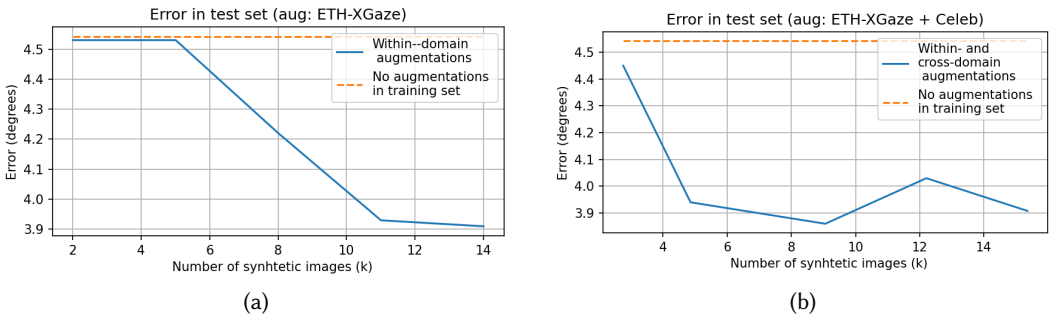(a)   (b)

Fig. 9.   Mean error of the gaze estimation DNN (test set) when trained with varying number of synthetic samples from the ETH-XGaze data distribution (a) and the ETH-XGaze and CelebAMask-HQ data distribution(b).

4.3   Gaze-aware Facial Image Generation

Facial image synthesis. The quality of image synthesis in generative models is evaluated using metrics such as the Fréchet Inception Distance (FID) [Heusel et al. 2017] and the Inception Score (IS) [Salimans et al. 2016]. Our models obtain a mean FID of 15.3, and a mean IS of 1.92. As far as we know, there are no generative models with input latent vector and gaze to compare. Consequently, we compare these metrics to the state-of-the-art methods for in-the-wild images' gaze editing.

DeepWarp and GazeGAN achieve a higher IS (2.89 and 3.10, respectively), while they also present a much higher FID (106.53 and 30.21, respectively). This suggests that our method can generate more realistic synthetic images. In addition, existing gaze redirection methods are typically limited to frontal gaze or gaze between two images, lacking fine-grained control.

As explained in Section 2, instead of training a generative model, we could synthesize facial images using non-DL approaches, such as graphics-based methods. However, the potential domain gap between real and synthetic images may limit their effectiveness in training. We generate 8,000 synthetic images using the 3D graphics-based method in [Wood et al. 2016] to evaluate gaze estimation accuracy when trained with these images compared to GC-GAN-generated ones. We train the gaze estimation DNN from [Zhang et al. 2020] with the same ETH-XGaze training set as the experiments in Tab. 1: only ETH-XGaze samples, in-domain and cross-domain synthetic images (as in Table 1), and additionally images generated using 3D graphics. [Wood et al. 2016] can produce images for a single eye, so we adjust our images by cropping the right half (left eye) to ensure equitable training conditions for the DNN. Fig. 10a shows the obtained results in the test set.

Training data	Error ↓
ETH-XGaze	4.88
ETH-XGaze + [Wood et al. 2016]	4.70
ETH-XGaze + GC-GAN (in-domain)	4.53
ETH-XGaze + GC-GAN (in-, cross-domain)	4.51

(a)   (b)

Fig. 10. Error in the ETH-XGaze test set (a) for single-eye crops with images generated by a non-DL method compared to GC-GAN. Samples of each training dataset (b), from left to right: ETH-XGaze, GC-GAN generated (in-domain), GC-GAN generated (cross-domain), and 3D graphics-based generated [Wood et al. 2016].

The error when training only with original samples (4.88°) is reduced when we add synthetic samples. However, this drop is higher with images generated with GC-GAN (4.51°). The results suggest that even though all synthetic images help, GC-GAN-generated ones are more effective. In addition, our approach learns from a specific data distribution and does not require a 3D environment, which can be time-consuming due to 3D asset design and rendering. Errors are higher compared to Table 1 (image containing both eyes), probably because of the reduced information.

Facial image editing. The GC-GAN allows selectively changing different face components, such as the nose, eyebrows, and face shape, while leaving the rest of the image unchanged by modifying the corresponding local latent vectors. Different semantic facial parts can be edited without training an additional model to manipulate the latent space of the model, which makes the model more interpretable for the user. In Fig. 11, we see an example of two synthetic images generated with the models trained in the first and second stage, given the same latent vector w and gaze direction θ.

Fig. 11. Images generated by models of stage 1 and stage 2 given the same gaze θ and latent vector w.

Next, we generate new augmented samples by editing a local latent vector w_k or the input gaze direction θ (Section 3.3) of the shown samples. Fig. 12 shows some examples generated by the model trained in the first and second stages, left and right images, respectively. We generate the images in the first row by modifying the latent vector of the eyebrows. We observe that augmented samples vary on the eyebrows but preserve the other face components unaltered, including the gaze direction. We generate the second-row images' by sampling new local latent vectors for the nose

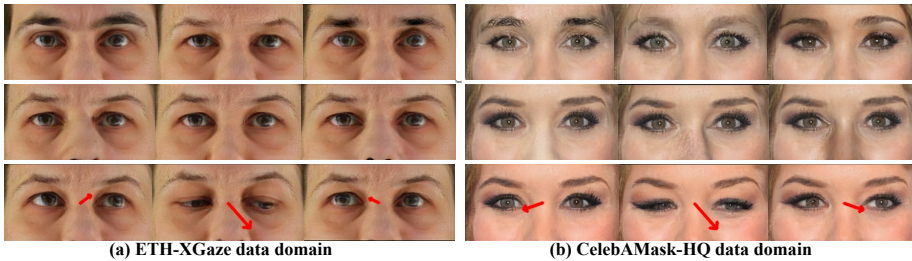(a) ETH-XGaze data domain   (b) CelebAMask-HQ data domain

Fig. 12.   Synthetic image augmentations by input latent or gaze vectors' modification in different domains.

generator and leaving the rest of the inputs the same as in Fig. 11. The images in the last row are generated by modifying the input gaze θ. Training datasets for DNNs should have enough samples and be varied and balanced. Consequently, having control over data generation is beneficial.

Gaze direction accuracy. We measure image quality by the IS and FID metrics, but these metrics do not consider the correctness of the subject's gaze in the synthetic image. This gaze should be consistent with the input target gaze to the model. In Section 4.2, we show that the gaze is accurate enough to help in DNN training. Next, we evaluate the gaze accuracy explicitly. We compute the gaze error between the input gaze to the model and the gaze estimated by a gaze estimation DNN on the synthetic sample. To minimize the possible errors of the gaze estimation DNN, we train the models in the Section 4.2 [Abdelrahman et al. 2022; Cheng and Lu 2022; Zhang et al. 2020] with the complete ETH-XGaze dataset and keep the two best checkpoints of each one. Then, we compute the mean of their output as the final gaze estimation for each tested image. We test all the synthetic images generated as image augmentations in Section 4.2 (in- and cross-domain). Fig 13a shows the obtained mean error in degrees for each of the four camera perspectives included in the dataset.

The mean error varies with perspective. More frontal perspectives (cameras 1 and 2) present lower error. Even if the obtained error is small, controlled-gaze image generation in lateral perspectives seems more challenging for the model. Fig. 13b shows some examples of synthetic images with arrows representing the gaze direction. Green represents the input target gaze to the GC-GAN, and red the estimated gaze by the DNN ensemble on the generated image. Despite some differences in gaze (from left to right, gaze errors: 3.33°, 1.78°, 4.7°, 8.84°), gaze directions are very similar.

Camera ID	1	2	3	4	Average
Mean error (°)	3.80	2.94	5.85	7.73	5.08
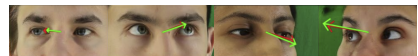
(a)   (b)

Fig. 13.   Gaze accuracy of synthetic images: (a) error between estimated and input gaze direction to the GC-GAN and (b) target gaze (green) for image generation and estimated gaze (red) on the generated images for different cameras (from left to right, camera 1 to 4).

Gaze direction accuracy in other benchmarks. To evaluate gaze accuracy under less controlled conditions compared to the ETH-X Gaze dataset, we extend the experiment to the MPIIFaceGaze dataset [Zhang et al. 2015]. Using the same training setup, we train the GC-GAN with data from 13 subjects in this dataset, achieving an FID of 17.66 and an IS of 1.84. As with the ETH-X Gaze augmented dataset (Section 4.2), we invert the original images to generate an augmented dataset, which we use to infer the gaze. For the gaze estimation DNN, we train the DNNs ensemble with all the images in the MPIIFaceGaze dataset and infer the gaze of generated 20,000 images. We obtain an average error of 5.29°, similar to Fig 13a. In this case, images are not captured in a laboratory and grouped by cameras. However, we obtain no significant difference between both dataset domains. Fig. 14a shows examples of a synthetic image in the MPIIFaceGaze data domain with different

augmentations (nose and gaze modifications). Fig. 14b compares input target gaze to the GC-GAN (green) and estimated gaze directions by the DNNs ensemble (red) for some generated samples.
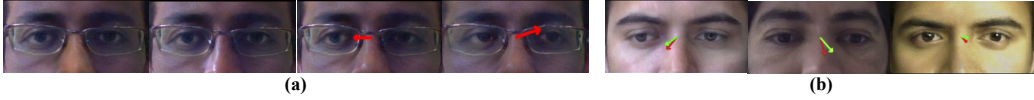

Fig. 14. Image augmentations by input latent or gaze vector modification in the MPIIFaceGaze data domain (a). Gaze accuracy in generated images (b): target gaze (green) for image generation and estimated gaze (red) on the generated images by the GC-GAN.

4.4   Ablation Studies: GC-GAN model training

What modules in Gaze-aware Compositional Generator should freeze during Stage-2? The generator has several combinations of modules to be frozen. We evaluate gaze transfer when training with different frozen modules. We qualitatively analyze gaze preservation by generating multiple pairs of images in both domains given a latent vector.

Table 2. Gaze preservation between domains when different components are frozen, Gaze-aware Local Generator (GLG) or Render Net (R), and whether mask constraint is applied.

ID	GLG	R	$Mask_c$	Gaze preservation
1	shape only	✗	✓	No
2	✓	✗	✓	No
3	✓	1	✓	No
4	✓	2	✓	Yes
5	✓	2	✗	No

In Table 2, we observe that freezing only the shape-oriented layers in the GLG is not enough, the weights of the whole GLG need to be preserved. The Render Net R also contains some information related to the gaze, as it is required to freeze the 2 initial residual blocks to transfer the gaze. It is also possible to freeze an additional block, but the image quality is worse. Regarding the segmentation mask constraint (Section 3.2), it is a necessary but not sufficient condition.

Components in Gaze-aware Generators. The GC-GAN design enables independent control of components and a disentangled latent space. We rely on the hypothesis that if some facial components (e.g., nose) are gaze-independent, their generator should learn without gaze conditioning. To demonstrate this, we retrain the GC-GAN with the ETH-XGaze dataset but modify the Local Generators so that all of them are gaze-conditioned and have the same architecture as a GLG.

We generate 5 random images with GC-GAN and the fully conditioned model. Then, we vary the input gaze of each sample to generate new images with 32 different gazes. We analyze how the images change, expecting only minor differences as only the eyes should vary. Differences are quantified using the pixel-wise mean absolute error between the original and redirected images. Separating gaze-related and unrelated generators reduces the error (0.93 versus 1.33). Fig. 15 shows a sample when varying the gaze. When conditioning all the generators (a), the gaze is entangled with the pose. The images generated with grouped generators (b) show a disentangled behavior: the eyes' appearance changes, but the rest of the image is unaltered.

4.5   Applications for gaze estimation

Our approach offers different potential applications to improve gaze estimation models or procedures. First, GC-GAN allows for combining labeled and unlabeled data sources, enabling training

Fig. 15. Synthetic images generated by varying the input gaze when all Local Generators are gaze-conditioned (a) and when generators are grouped as in GC-GAN (b).

with unannotated in-the-wild images. This is particularly valuable given gaze annotations' laborious and costly nature and the importance of data quantity and variability in training DNNs. Our method can also be used to enhance the predictability of gaze estimation models by extensive model evaluation under different conditions. The synthetic images generated with GC-GAN could be used to evaluate a model's robustness against image variations and detect potential limitations or biases when using specific input data (e.g., different facial attributes, gaze distributions, or conditions, like glasses). This analysis could be used to improve models, predict their performance under different circumstances, and make their behavior more trustworthy. Additionally, our approach streamlines calibration by generating new samples within a domain and subject using a smaller image set. This enhances capturing efficiency and may improve model personalization (e.g., specific subject).

Privacy and ethics. The proposed method may inherit the negative societal impacts of existing deep generative models, such as improperly generating fake data. However, we believe our proposed method can be beneficial to build more robust gaze estimation models that guarantee universal human-computer interaction systems and have a smaller bias to certain people groups for correct behavior. Different approaches might be used to account for possible model biases or to avoid improper use. Variational autoencoders might be used to identify possible dataset bias [Amini et al. 2019], which synthetic images could later mitigate. Another possibility to promote responsible artificial intelligence is the data and model cards [Mitchell et al. 2019; Pushkarna et al. 2022]. Model cards are designed to accompany models to clarify their intended use, provide a model evaluation in various conditions (e.g., race, geographic location), or register the provenance of training datasets.

5 CONCLUSION

In this paper, we have presented a novel method to generate annotated gaze data by combining the strengths of different labeled and unlabeled data sources. First, we train a gaze-aware compositional GAN that is able to generate realistic synthetic images with specific gaze directions in a labeled data domain. Then, we transfer this model to an unlabeled data domain to exploit the variance these data provide. Experiments have shown that our method can be used to augment existing annotated data for gaze estimation DNN training by generating within-domain and cross-domain augmentations that boost the DNN accuracy. The presented method does not require 3D virtual environments to generate synthetic data for DNN training and generates highly realistic samples, as the distribution of the generated images is learned from the training domain. We have also shown other applications of our work to facial image editing and gaze redirection. We believe the current work presents an exciting direction to leverage existing data and improve state-of-the-art computer vision models from a data-centric perspective. In future work, we will explore incorporating the camera pose for fine-grained controlled image generation.

ACKNOWLEDGMENTS

REFERENCES

Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. 2022. L2CS-Net: fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339* (2022).

Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 289–295.

Nuri Murat Arar, Hua Gao, and Jean-Philippe Thiran. 2016. A regression-based user calibration framework for real-time gaze estimation. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 12 (2016), 2623–2638.

Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. 2019. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047* (2019).

Sagie Benaim and Lior Wolf. 2018. One-shot unsupervised cross domain translation. *advances in neural information processing systems* 31 (2018).

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).

Jingjing Chen, Jichao Zhang, Enver Sangineto, Tao Chen, Jiayuan Fan, and Nicu Sebe. 2021. Coarse-to-fine gaze redirection with numerical and pictorial guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3665–3674.

Yihua Cheng and Feng Lu. 2022. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 3341–3347.

Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. 2021. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668* (2021).

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine* 35, 1 (2018), 53–65.

Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. 2020. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5154–5163.

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.

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

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

Takahiro Ishikawa. 2004. Passive driver gaze tracking with active appearance models. (2004).

Ali Jahanian, Lucy Chai, and Phillip Isola. 2019. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171* (2019).

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34 (2021), 852–863.

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

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. 2019. Real-time facial surface geometry from monocular video on mobile GPUs. *arXiv preprint arXiv:1907.06724* (2019).

Harsimran Kaur and Roberto Manduchi. 2020. Eyegan: Gaze-preserving, mask-mediated eye image synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 310–319.

Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6912–6921.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5549–5558.

Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C Buehler, Otmar Hilliges, and Thabo Beeler. 2022. EyeNeRF: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–16.

David Masko. 2017. Calibration in eye tracking using transfer learning.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which training methods for GANs do actually converge?. In *International conference on machine learning*. PMLR, 3481–3490.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65 (2021), 99–106.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

Sergey I Nikolenko. 2021. *Synthetic data for deep learning*. Vol. 174. Springer.

Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. 2019. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9368–9377.

Sonia Porta, Benoit Bossavit, Rafael Cabeza, Andoni Larumbe-Bergera, Gonzalo Garde, and Arantxa Villanueva. 2019. U2Eyes: A binocular dataset for eye tracking and gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1776–1826.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

Matan Sela, Pingmei Xu, Junfeng He, Vidhya Navalpakkam, and Dmitry Lagun. 2017. Gazegan-unpaired adversarial image generation for gaze estimation. *arXiv preprint arXiv:1711.09767* (2017).

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9243–9252.

Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. 2022. SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11254–11264.

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

Yezhi Shu, Ran Yi, Mengfei Xia, Zipeng Ye, Wang Zhao, Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. 2021. Gan-based multi-style photo cartoonization. *IEEE Transactions on Visualization and Computer Graphics* 28, 10 (2021), 3376–3390.

Neelabh Sinha, Michal Balazia, and François Bremond. 2021. FLAME: Facial Landmark Heatmap Activated Multimodal Gaze Estimation. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–8.

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

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Wensen Feng. 2020. Controllable continuous gaze redirection. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1782–1790.

Jichao Zhang, Meng Sun, Jingjing Chen, Hao Tang, Yan Yan, Xueying Qin, and Nicu Sebe. 2019. Gazecorrection: Self-guided eye manipulation in the wild using self-supervised generative adversarial networks. *arXiv preprint arXiv:1906.00805* (2019).

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

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. 2020. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*. Springer, 365–381.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4511–4520.