Title

Subject Guided Eye Image Synthesis with Application to Gaze Redirection.

ISBN

978-1-6654-0477-8

Authors

Kaur, Harsimran

Manduchi, Roberto

Publication Date

2021

DOI

10.1109/WACV48630.2021.00006

Peer reviewed

Subject Guided Eye Image Synthesis with Application to Gaze Redirection

Harsimran Kaur Roberto Manduchi
University of California, Santa Cruz
{hkaur14, manduchi}@ucsc.edu

Abstract

We propose a method for synthesizing eye images from segmentation masks with a desired style. The style encompasses attributes such as skin color, texture, iris color, and personal identity. Our approach generates an eye image that is consistent with a given segmentation mask and has the attributes of the input style image. We apply our method to data augmentation as well as to gaze redirection. The previous techniques of synthesizing real eye images from synthetic eye images for data augmentation lacked control over the generated attributes. We demonstrate the effectiveness of the proposed method in synthesizing realistic eye images with given characteristics corresponding to the synthetic labels for data augmentation, which is further useful for various tasks such as gaze estimation, eye image segmentation, pupil detection, etc. We also show how our approach can be applied to gaze redirection using only synthetic gaze labels, improving the previous state of the art results. The main contributions of our paper are i) a novel approach for Style-Based eye image generation from segmentation mask; ii) the use of this approach for gaze-redirection without the need for gaze annotated real eye images

1. Introduction

There is intense recent interest in the synthesis of realistic images of human faces with a prescribed gaze direction. While model-based computer graphics methods have long been used for this purpose (e.g.[3]), they typically generate images that do not look "natural". In this paper we address the specific problem of *gaze redirection*: given an image of a person's face, we want to generate a new image that is identical to the first one, except for this person's gaze, which should be consistent with a certain direction.

Gaze redirection finds multiple applications, including videoconferencing (making people appear as if they were looking at the camera [20]), photo correction [32], and video editing. Gaze redirection may also be a useful tool for generating data sets that can be used to train appearance-base gaze estimation algorithms [26], [37], [38]. Indeed, acquiring large quantities of annotated data (with gaze direction for each image) can be time-consuming and prone to error, since accurate measurement of gaze direction is difficult to achieve in practice. Some recent works used training images rendered by means of computer graphics methods [33], [34]. While this approach has the advantage of providing accurate gaze and eye feature annotations, the rendered images are far from real. As a consequence, models trained on these images may not do well when applied to real-world eye images. To overcome this problem, some authors have proposed methods for synthesizing real-world eye images from synthetic eye images while maintaining annotations (e.g, gaze direction; [28], [21], [15].) While effective, these eye image synthesis methods do not give the user much control over the generated eye features - skin color, texture, iris color, and eye shape. Our method builds on this previous work, but operates on a specific desired eye image instance, rather than on a generic model.

We cast the task of gaze redirection as one of image synthesis with a pre-determined style. By *style* we mean, for example, the appearance of a certain person's face under a certain illumination. In this context, the *content* to be manipulated is gaze direction. Following Kaur et al. [15], we use ternary *segmentation masks* to characterize gaze direction. Masks act as style-independent proxies for gaze. Kaur et al. [15] introduced an algorithm (EyeGAN) that takes a synthetic mask for a prescribed gaze direction as input, and generates an image under content (gaze direction) consistent with the input mask and some random style. In this work, we push this idea forth, and introduce a new cyclic mechanism to ensure consistency of both style and content of the generated image. In addition, we introduce an algorithm that redirects one's gaze without relying on model-based synthetic mask generation. A ternary mask is extracted from the input image, and redirected (using a trained network) to the desired gaze direction. This new mask is then used to control style-preserving gaze redirection. Remarkably, our algorithms *do not* require gaze annotated real-world images for training.

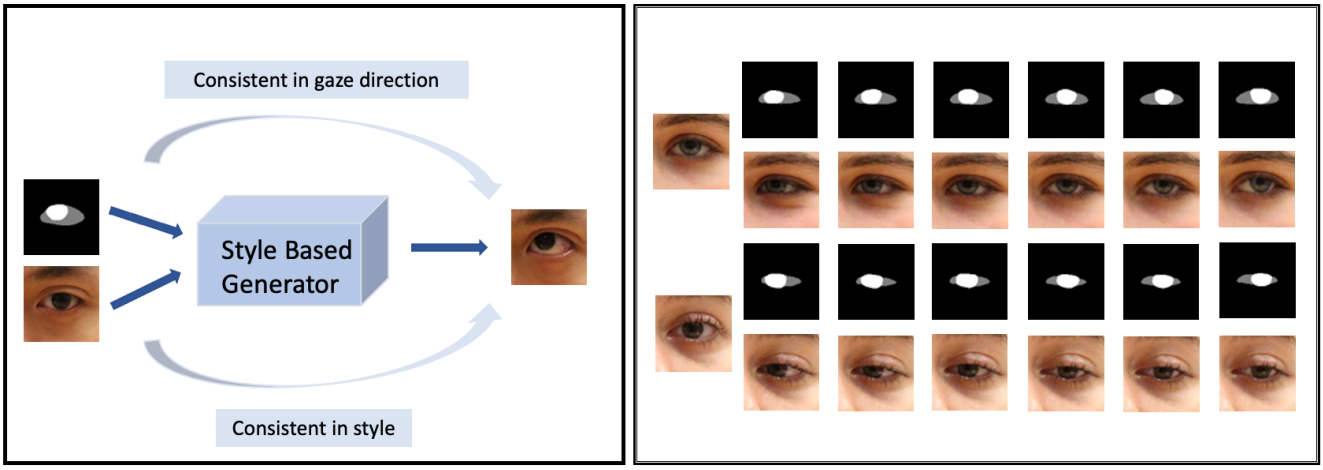


Figure 1: Overview of our Style-Based Eye Image Generation. The generator receives in input a segmentation mask and a style image. It synthesizes an image which is consistent in gaze with the segmentation mask, with generated features similar to the style image.

2. Related Work

Generative Adversarial Networks. Generative Adversarial Networks (GANs) [7] have gained tremendous success for image generation tasks. The use of adversarial loss using a discriminator network has been shown to improve the quality of generated images compared to using traditional losses (e.g. least squares). In this work, we include GAN adversarial loss for eye image synthesis. Conditional GANs [25] are widely used in image–to–image translation tasks. One standard approach is to train an image translation model using paired data [12], [31], e.g. pairing an RGB image with its associated segmentation mask or edge map. Then, at run time, only one of the two images (e.g., edge map) is input to the system, which generates the associated image (e.g. RGB). When paired data is not available, unsupervised methods can be used [39], [28], [16] ,[23]. While SimGAN [28] and CycleGAN [21] translate synthetic eye images into real world images, EyeGAN [15] starts from ternary eye segmentation masks. EyeGAN is trained using pix2pix [12] on image/segmentation input pairs, where the segmentation mask is extracted from the input image. The segmentation network is trained in tandem with the eye synthesis network, using auxiliary synthetically generated image/segmentation pairs.

Style-Based Image Generation. The methods cited above generate new images in the style of the images used in the training data. While this may be acceptable for purposes such as producing a realistic data set, tasks such as gaze redirection call for precise control of the style at run time. Stated differently, a gaze redirection algorithm must ensure that the generated image is consistent with the style of the input image – it must preserve the features that characterize the appearance of the person in the image.

Image style can be modeled as a learned distribution using variational auto-encoders [18] At inference time, one can sample the style from the learned distribution to generate the image [40] [27]. A method for deterministic generation of images with a specific style using a gram-matrix based style loss was proposed in [5]. The work in [11] showed that style could be transferred from input image to synthesized image using adaptive instance normalization. StyleGAN [14] used GAN based generator with adaptive instance normalization to synthesize novel human face images. The work in [1] also used adaptive normalization with a SPADE [27] generator for synthesizing eye images from segmentation masks consistent with the input style. Wang et al. [30] used style consistent as well style inconsistent pairs as an input to a discriminator, in order to impose the input style in the output image. In our work, we use a cyclic loss to enforce style.

Gaze Manipulation. Earlier work on gaze manipulation focused correcting gaze direction such that the person appears to be looking at the camera, which is very desirable for applications such as videoconferencing. Some of the proposed approaches required specialized 3D data capture hardware to synthesize novel views of face and eyes [20], [6], [2]. Gaze redirection is a more generic task of manipulating the gaze to any arbitrary direction. Monocular image-based gaze redirection can be achieved by learning a warping flow field between images with a known correction angle. This flow field can be computed using Random forests [19] or deep networks [4]. In [35], a flow field network is trained on synthetic eye images for gaze redirection, and domain adaptation is applied from synthetic to realistic

eye images. This work was primarily focused on improving user-specific gaze estimation by using few-shot learning. Wood et al. [32] used a 3D morphable eye model for gaze manipulation. The work in [9] used GANs to synthesize eye images with redirected gaze using a specific reconstruction loss. Our work is similar to [9], however, rather than guiding synthesis by a gaze angle vector, we use segmentation masks. This mitigates the need for obtaining annotated gaze data for real-world eye images.

3. Method

3.1. Style–Based Eye Image Synthesis

The goal of this module is to generate a realistic image with a certain style and a prescribed gaze direction. Style is guided by means of an eye image from a domain $\mathcal{E}$. This image is given in input to the network, along with a ternary segmentation mask (from the mask domain $\mathcal{M}$), which characterizes the prescribed gaze direction. Thus, our system is a mapping G from $\mathcal{M} \times \mathcal{E}$ to $\mathcal{E}$. Fig 1 shows an overview of our style–based eye image synthesis.

Let $E_s^{g_1} \in \mathcal{E}$ be the input eye image, where the subscript $_s$ indicates the style, and the superscript ^{g_1} indicates the gaze direction in the image. Note that "style" is not a directly measurable quantity – it expresses the appearance of the eye image, in terms, for example, of iris color, skin color, skin texture, illumination. Gaze is quantifiable, but we don't need to know, nor make use of, the gaze direction g_1. The goal of the generator G is to synthesize an eye image $E_s^{g_2*} \in \mathcal{E}$ with same style s as the input, but with gaze direction g_2 (the superscript * indicates that this is a synthesized image.) The algorithm uses a synthetically generated ternary mask, M^{g_2} as a proxy for the prescribed direction g_2.

The network is trained using samples consisting of four images each: $\{E_s^{g_1}, M(E_s^{g_1}), E_s^{g_2}, M(E_s^{g_2})\}$. Here, $E_s^{g_1}$ and $E_s^{g_2}$ are images of the same individual with different gaze directions. $M(\cdot)$ is a function that extracts a ternary mask from an eye image [15].

Synthesis Loss. The *synthesis loss* term ensures that the generated image, $E_s^{g_2*}$, is similar to the desired one. For this purpose, we use a perceptual loss function [13], [9]:

$$\mathcal{L}_{syn}(G) = \sum_j w_j * \frac{1}{N_j} ||f_j(E_s^{g_2*}) - f_j(E_s^{g_2})||_2^2 \quad (1)$$

$$= \sum_j w_j * \frac{1}{N_j} ||f_j(G(M^{g_2}, E_s^{g_1})) - f_j(E_s^{g_2})||_2^2 \quad (2)$$

Here, $f_j(\cdot)$ is the feature map of size $N_j = C_j \times H_j \times W_j$, extracted from j^{th} convolutional layer in the VGG-16 network, pre-trained on the ImageNet data set.

Re-synthesis Loss. As an additional device to ensure style consistency between the input $E_s^{g_1}$ and the synthesized image $E_s^{g_2*}$, the latter is taken as input to the generator during training along with the segmentation mask M^{g_1} from the input image, to generate a new image $E_s^{g_1*} = G(M^{g_1}, E_s^{g_2*})$. A second loss component is added as follows:

$$\mathcal{L}_{resyn}(G) = ||E_s^{g_1*} - E_s^{g_1}||_1 \quad (3)$$

$$= ||G(M^{g_1}, G(M^{g_2}, E_s^{g_1})) - E_s^{g_1}||_1 \quad (4)$$

This idea borrows from the CycleGAN scheme [21], with the difference that CycleGAN maps an image from a domain to a different domain and back, while we map an image to the same domain (and back), but with an additional "content guidance" image (the ternary segmentation representing gaze direction.)

Adversarial Loss. In order to improve the quality of image generation, we also consider an adversarial loss [7]. A discriminator network $D(\cdot)$ is shown an image, either synthesized ($E_s^{g_2*}$) or real ($E_s^{g_2}$), and is tasked with determining whether the input image is real or synthesized. This loss term penalizes the discriminator when the determination is incorrect, and the generator when it is correct:

$$\mathcal{L}_{adv}(G, D) = \log(D(E_s^{g_2})) + \log(1 - D(E_s^{g_2*})) \quad (5)$$

$$= \log(D(E_s^{g_2})) + \log(1 - D(G(M^{g_2}, E_s^{g_1}))) \quad (6)$$

Along with the adversarial loss, we consider a discriminator-based feature matching loss [31]. Features are extracted from intermediate layers of the discriminator network for both a real ($E_s^{g_2}$) and synthesized ($E_s^{g_2*}$) image. This loss penalizes discrepancy between the features for the two images.

$$\mathcal{L}_{feat}(G) = \sum_j \frac{1}{N_j} ||D_j(E_s^{g_2*}) - D_j(E_s^{g_2})||_2^2 \quad (7)$$

$$= \sum_j \frac{1}{N_j} ||D_j(G(M^{g_2}, E_s^{g_1}) - D_j(E_s^{g_2})||_2^2 \quad (8)$$

Here D_j represents the feature map extracted from the j^{th} layer of the discriminator network. The size of the feature map is given by $N_j = C_j \times H_j \times W_j$.

Overall Objective The overall objective is given by:

$$G^*, D^* = \min_G \max_D (\mathcal{L}_{adv}(G, D) + \lambda_1 \mathcal{L}_{feat}(G) + \lambda_2 \mathcal{L}_{syn}(G) + \lambda_3 \mathcal{L}_{resyn}(G)) \quad (9)$$

3.2. Gaze Redirection via Mask Synthesis

In the algorithm described above, gaze redirection is guided by a ternary mask M^{g_2}, which describes the desired

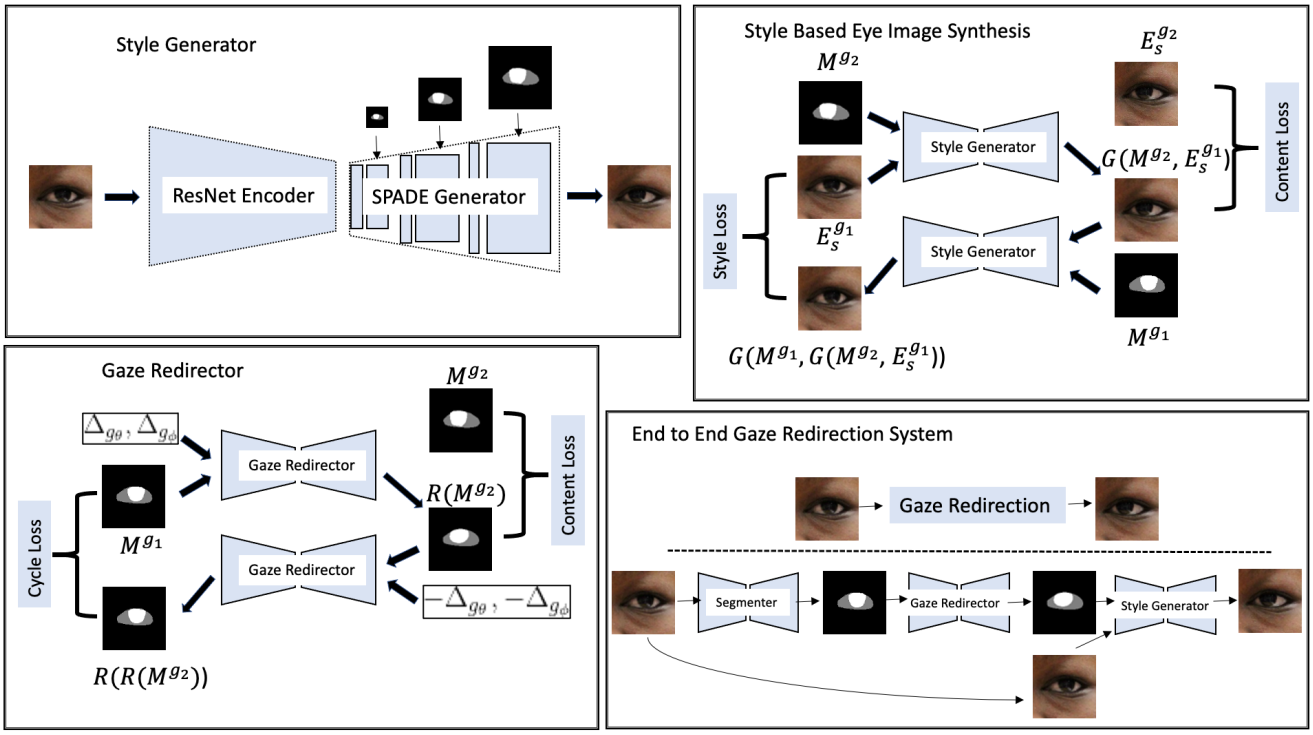


Figure 2: Top Left: SPADE [27] based Generator architecture. The ResNet Encoder encodes the style image. The style encoding is input to the SPADE generator which also receives segmentation mask input at different scales. Top Right: Our Style-Based Eye Image Synthesis flow. The generator synthesizes the eye image from the mask and style image. During training, the generated image is fed back to the generator along with mask corresponding to the style image. Bottom Left: Gaze Redirection model trained on the segmentation masks. Bottom Right: End to End Gaze redirection system involving segmentation mask generator, gaze redirector and style generator.

gaze direction. This mask would normally be obtained using a computer graphics tool, such as UnityEyes [33]. However, this mask may not be perfectly suited to the considered "style". Due to the variability of facial features, a synthetic mask can be only an approximation of the actual segmentation obtained from a real image. For this reason, a synthetic mask may only be a sub-optimal solution for guiding gaze redirection.

We address this problem with an algorithm that builds on our Style–Based Eye Image Synthesis approach, but that synthesizes the content-guiding ternary mask from the segmentation of the input image. The hope is that this mask may represent a better proxy for our gaze redirection algorithm. Mask synthesis is the job of a *mask redirection* network, which takes in input a ternary segmentation of the input image $M(E_s^{g_1})$, along with the the prescribed variation of gaze angle $(\Delta_\phi, \Delta_\theta) = (g_2^\phi, g_2^\theta) - (g_1^\phi, g_1^\theta)$ [9] [35], to produce a redirected mask $M^{g_2 *}$.

The mask redirection network R is trained with pairs of segmentation masks $(M(E_s^{g_1}), M(E_s^{g_2}))$ from images with known gaze directions (g_1, g_2). A loss function is defined as the sum of two terms: a mask-synthesis loss and a mask-resynthesis loss.

Mask-Synthesis Loss. This is a forward content loss between the gaze redirected mask and the target mask.

$$\mathcal{L}_{m-syn}(R) = \text{CE}[R(M(E_s^{g_1}), (\Delta_\phi, \Delta_\theta)) - M(E_s^{g_2})]$$

Mask-Resynthesis Loss. The redirected mask is fed back to the network with negative gaze direction variation, with the goal to reconstruct the input mask:

$$\mathcal{L}_{m-resyn}(R) = \text{CE}[R(R(M(E_s^{g_1}), (\Delta_\phi, \Delta_\theta)), (-\Delta_\phi, -\Delta_\theta)) - M(E_s^{g_1})]$$

In these equations, CE$[\cdot]$ represents the cross-entropy function.

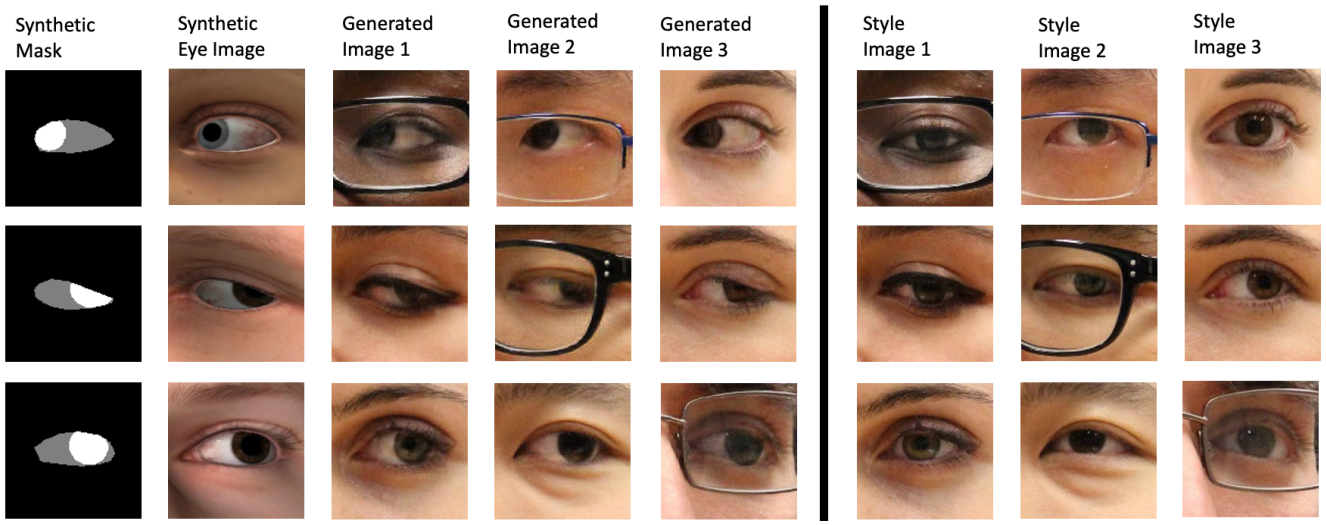


Figure 3: Eye Image Synthesis from Unity Masks. First and second column: masks from UnityEyes and corresponding synthetic eye images. Columns 3-5: the generated images from our Style-Based generator for the mask in Column 1. The input style images are shown in last three columns.

4. Experiments

4.1. Style-Based Eye Image Synthesis

Data Set. We trained our Style–Based Eye Image Synthesis network on the Columbia Gaze data set [29]. This data set contains facial images taken of 56 subjects under constrained settings. For each subject, gaze data was collected at 5 horizontal head poses $[0°, \pm 15°, \pm 30°]$. For each head pose, 21 gaze directions (7 horizontal: $\phi \in [0°, \pm 5°, \pm 10°, \pm 15°]$ and 3 vertical $\theta \in [0°, \pm 10°]$) were recorded. We cropped the left and right eye patches from the facial images and resized them to 64×64 as described in [9]. The right eye images were flipped to look like left eye images. Each style category consists of images for one subject with different gaze directions with one head pose and one side (left/ right). In practice, a total of $56 \cdot 5 \cdot 2 = 560$ styles were considered (56 subjects with 5 different head poses and 2 sides). Each style has 21 images (one per gaze direction). We used the first 50 subjects for training and the remaining 6 subjects for testing.

We used EyeGAN [15] to extract segmentation masks for all of the eye images. This method trains a fully convolutional neural net [24] to extract segmentation masks without manual annotations. It uses a set of real eye images along with synthetic eye masks (from the UnityEyes [33] tool.) It alternates training of a segmenter network to extract masks from real eye images, with training of a generator network that synthesizes natural looking eye images from the synthetic masks. Since the mask generated by EyeGAN have a smaller size 48×32, we zero-padded them to fill a 64×64 area.

Implementation Details Our eye image synthesis generator is implemented as an encoder-decoder network. In particular, we use a ResNet [8] encoder with three residual blocks, and a SPADE [27] decoder for generating images guided by segmentation masks (see Fig. 2). In SPADE, the segmentation mask is fed at each block at different scales. Our SPADE decoder consists of four SPADEResNet blocks. Training is performed using the same multi-scale discriminator as in pix2pixHD [31] and SPADE [27] with hinge loss [22], [27]. We used Adam [17] optimizer with $\beta_1 = 0$ and $\beta_2 = 0.9$. Learning rate was set to 0.005 with λ_1, λ_2 and λ_3 as 10, 20 and 20 respectively.

4.2. Gaze Redirection via Mask Synthesis

Data Set The mask redirection network was trained with synthetic masks corresponding to the eye images from 10 different subjects (different eye parameters) generated for 21 gaze directions in frontal head pose using UnityEyes [33] tool. For each gaze direction, we trained the network to redirect the mask to 20 other gaze directions.

Implementation Details The mask redirection network receives in input a segmentation mask and a gaze direction variation vector. The input segmentation mask is passed through a network with three convolutional layers, followed by five residual blocks and then by three upsampling + convolutional layers. The gaze direction variation vector is input to a Multi-Layer Perceptron, then concatenated with output of the convolutional layers for the input mask, before the residual blocks. The network is trained to minimize per-

Table 1: Comparison of eye image synthesis algorithms using FID score (lower the better) and mIOU (higher the better).

Algorithm	FID	mIoU
EyeGAN [15]	83.9	**0.93**
CycleGAN [39], [21]	39.5	0.61
SimGAN [28]	53.7	0.66
Ours	**8.5**	0.72

pixel cross-entropy loss between the generated mask and the ground truth mask from UnitiEyes. We used Adam [17] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was set to 0.01.

The output of the mask redirection network is used as input mask in the Style-Based Eye Image Synthesis (Sec. 3.1).

5. Results

Style-Based Eye Image Synthesis. In this case, eye images are synthesized with guidance from the UnityEyes segmentation masks. We first compare our Style-Based Eye Image Synthesis with three other eye image synthesis algorithms: SimGAN [28], CycleGAN [39], [21] and EyeGAN [15]. These algorithms synthesize an image with the style of the data set on which they are trained (in this case, the Columbia Gaze data set.) In the case of of SimGAN and CycleGAN, the input to the algorithm is a synthetic eye image from UnityEye. EyeGAN takes in input a segmentation mask, also from UnityEye.

Two metrics were considered for comparison: (1) Fréchet Inception Distance (FID) [10]; and (2) mean IoU (mIoU). FID is a metric used to measure similarity between two data sets (in our case, the output of the algorithms and the Columbia Gaze data set.) It captures both the perceptual similarity between generated and real images, and the diversity of generated images (similar data sets have low FID values). The mean IOU is computed between the segmentation of the output, and the mask corresponding to the input image or the mask itself in our case. This segmentation is obtained by a FCN [24] trained on masks corresponding to UnityEyes images and the corresponding eye images generated using EyeGAN since it has been shown to preserve semantic consistency of the generated images. A larger value of mIoU indicates good semantic consistency between the source and the generated image.

Table 1 presents quantitative results in terms of the FID and mIoU metrics. Our technique achieves the smallest value of FID, and the second highest value of mIOU among the algorithms considered. Samples of eye images generated by our method, along with the "style images" and synthetic masks used in input, are shown in Figure 3.

We also compared our algorithm with two other techniques for style-based synthesis: Seg2Eye [1] and Pix2PixSC [30] on Columbia eye data set [29]. Seg2Eye uses the SPADE [27] architecture with adaptive-instance normalization layers for style transfer. We trained Seg2Eye with a single style image per data point. Pix2PixSC uses style consistency adversarial loss with Pix2PixHD [31] as a base architecture.

Table 2: Comparison of Style-Based eye image synthesis using LPIPS metric (lower the better).

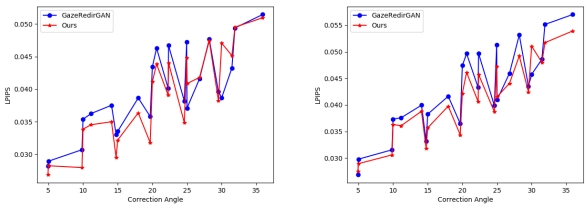
(a) LPIPS score on the test data with **supervised** training.

Algorithm	LPIPS
Pix2PixSC [30]	0.104
Seg2Eye [1]	0.077
Ours w/o reconst	0.035
Ours	**0.033**

(b) LPIPS score on the test data with **unsupervised** training.

Algorithm	LPIPS
Pix2PixSC [30]	0.125
Seg2Eye [1]	0.124
Ours w/o reconst	0.078
Ours	**0.044**

We used the segmentation mask corresponding to test subject images and generated the style image corresponding to the masks. The generated images are compared with the ground truth images using LPIPS [36] metric. We trained our network with baseline under two kinds of settings, supervised and unsupervised. In supervised setting, the ground truth image corresponding to the input mask has the same style as the input style image. In unsupervised setting the ground truth image has the different style as the input style image. We observed that, with supervised training, SPADE generator architecture with ResNet encoder in itself is good enough to generate Style-Based images, without the need for our re-synthesis loss. However, our re-synthesis loss proved very useful in case of unsupervised training. We show the LPIPS metric results for the two baselines and our method with and without reconstruction in Table 2, for both supervised and unsupervised training. We also show the qualitative results in Figure 5 for supervised training and Figure 6 for unsupervised training.



(a) Full Dataset. (b) Reduced Dataset.

Figure 4: LPIPS vs Correction Angle: Quantitative comparison of gaze redirection results for frontal head pose.

Gaze Redirection via Mask Synthesis. We compared our algorithm for gaze redirection based on mask synthe-

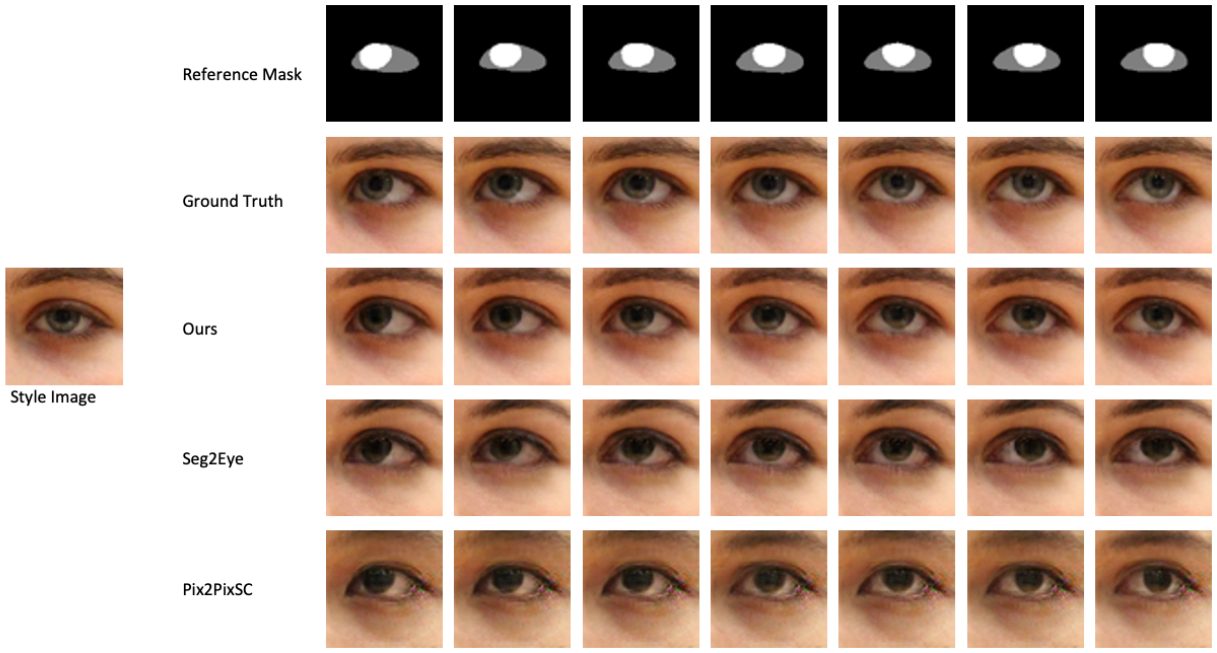


Figure 5: Qualitative comparison under supervised setting. We show the results corresponding to different segmentation masks for a test style image. The ground truth images are shown along with the synthesized images for baseline methods.

Figure 6: Qualitative comparison under unsupervised setting. We show the results corresponding to different segmentation masks for a test style image. The ground truth images are shown along with the synthesized images for baseline methods.

sis against the method described in [9] (which we dubbed "GazeRedirGAN",) which is shown to give state of the art results.

The LPIPS [36] metric was used for comparing generated and ground truth images. The mean LPIPS score is calculated with respect to the correction angle [9], which is the angular difference between target gaze direction and source gaze direction. As shown in Figure 4a, our method per-

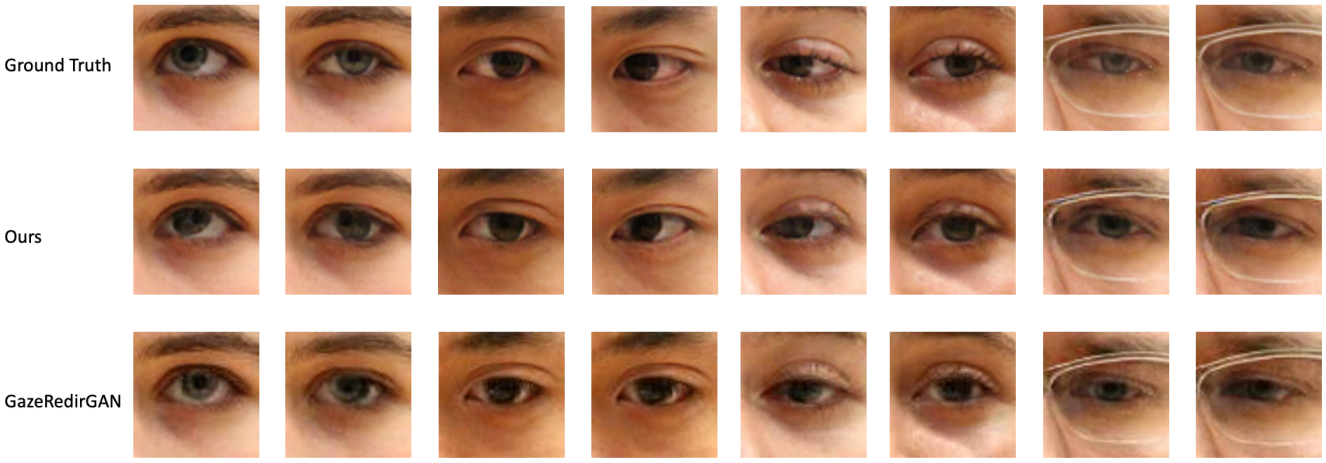


Figure 7: Qualitative Comparison of gaze redirection results with models trained on reduced data set.

forms slightly better than GazeRedirGAN, even though we only used the gaze labels corresponding to synthetic masks for redirection.

In another experiment, we removed eye images corresponding to horizontal angles $[\pm 10°, \pm 15°]$. We trained both our Style-Based Eye Image Synthesis algorithm and GazeRedirGAN on this reduced (Columbia) data set, and tested on the gaze directions unseen in the training set. The synthetics masks corresponding to the removed gaze directions were used to train the gaze redirection network. As shown in Figure 4b our gaze redirection produces lower LPIPS values. We also show qualitative comparison in Figure 7.

6. Conclusion

We proposed a new method for generating realistic eye images with a prescribed gaze direction. This algorithm takes in input a "style image" as well as a ternary segmentation mask, representing the desired gaze direction. A cyclic training algorithm ensures that the generated image has the desired gaze direction, and that it is in the style of the input image.

We also show how we can use this style synthesis for gaze redirection. Importantly, this algorithm does not require annotation of gaze angle in the training data. Instead, it uses ternary segmentation of the training images, which is much easier to obtain. The gaze labels are required corresponding to the ternary mask which can be generated synthetically.

7. Acknowledgment

Research reported in this publication was supported by the National Eye Institute of the National Institutes of Health under award number R01EY030952-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Marcel C. Buehler, Seonwook Park, Shalini De Mello, Xucong Zhang, and Otmar Hilliges. Content-consistent generation of realistic eyes with style. In *International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[2] Criminisi, Shotton, Blake, and Torr. Gaze manipulation for one-to-one teleconferencing. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 191–198 vol.1, 2003.

[3] Zhigang Deng, John P Lewis, and Ulrich Neumann. Automated eye motion using texture synthesis. *IEEE Computer Graphics and Applications*, 25(2):24–30, 2005.

[4] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 311–326, Cham, 2016.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[6] D. Giger, J. Bazin, C. Kuster, T. Popa, and M. Gross. Gaze correction with a single webcam. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2014.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Z. He, A. Spurr, X. Zhang, and O. Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6931–6940, 2019.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30*, pages 6626–6637. 2017.

[11] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017.

[12] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[13] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, pages 694–711, 2016.

[14] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[15] H. Kaur and R. Manduchi. Eyegan: Gaze-preserving, mask-mediated eye image synthesis. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[16] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1857–1865, 2017.

[17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[19] D. Kononenko and V. Lempitsky. Learning to look up: Realtime monocular gaze correction using machine learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4667–4675, 2015.

[20] Claudia Kuster, Tiberiu Popa, Jean-Charles Bazin, Craig Gotsman, and Markus Gross. Gaze correction for home video conferencing. *ACM Trans. Graph. (Proc. of ACM SIGGRAPH ASIA)*, 2012.

[21] Kangwook Lee, Hoon Kim, and Changho Suh. Simulated+unsupervised learning with adaptive data generation and bidirectional mappings. In *International Conference on Learning Representations*, 2018.

[22] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

[23] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems 30*, pages 700–708. 2017.

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[25] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[26] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research Applications*, 2018.

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[28] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2242–2251, 2017.

[29] B.A. Smith, Q. Yin, S.K. Feiner, and S.K. Nayar. Gaze Locking: Passive Eye Contact Detection for Human Object Interaction. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 271–280, 2013.

[30] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter. M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[31] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[32] Erroll Wood, Tadas Baltrusaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Gazedirector: Fully articulated eye gaze redirection in video. *Comput. Graph. Forum*, pages 217–225, 2018.

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

[34] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV 2015)*, 2015.

[35] Y. Yu, G. Liu, and J. Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11929–11938, 2019.

[36] 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.

[37] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 4511–4520, 2015.

[38] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2299–2308, 2017.

[39] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[40] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems 30*, pages 465–476. 2017.