EyeGAN: Gaze–Preserving, Mask–Mediated Eye Image Synthesis

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

Abstract

Automatic synthesis of realistic eye images with prescribed gaze direction is important for multiple application domains. We introduce EyeGAN, an algorithm to generate eye images in the style of a desired target domain, that inherit annotations available in images from a source domain. EyeGAN takes in input ternary masks, which are used as domain-independent proxies for gaze direction. We evaluate EyeGAN against competing eye image synthesis algorithms by measuring a specific gaze consistency index. In addition, we present results from multiple experiments (involving eye region segmentation, pupil localization, and gaze direction estimation) showing that the use of EyeGAN-generated images with inherited annotations for network training leads to superior performances compared to other domain transfer algorithms.

1. Introduction

We are interested in generating realistic images of human eyes with a prescribed gaze direction. A direct practical applications of this technology is gaze redirection for teleconferencing [4]. A more indirect application is the creation of data sets for the training of image-based gaze tracking algorithms. These systems require large amounts of images with specific annotations. While some annotation types (e.g., the location of the pupil center) can be easily obtained via manual labeling, others are more challenging. For example, in order to determine the gaze direction of people visible in the images, data sets are often built by asking human subjects to look at a certain point on a screen [42] or at a calibrated location (such as an object [7]). Then, gaze direction annotations are extrapolated from geometric reasoning, such as by drawing a line from the location on the screen been fixated to the viewer's pupil, whose location in 3-D is assumed known. This is a relatively complex and error-prone procedure. Other features that cannot be obtained by manual labeling (because not observable) include the center of rotation of the subject's eyeball, which is needed to train model-based gaze tracking algorithms [3, 37].

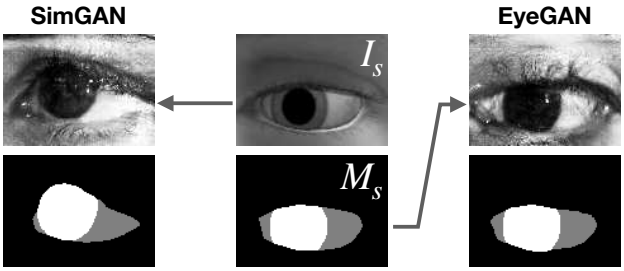

Figure 1. Center: Image–mask (I_s, M_s) pair synthesized by UnityEyes [39]. Left: The image generated by SimGAN [33] using I_s as input. Right: The image generated by our EyeGAN system using M_s as input. Both generated images are shown with the associated mask, computed by the segmenter trained with EyeGAN.

Several methods have been proposed and demonstrated for the generation of realistic eye images, with generative adversarial networks (GAN [9]) arguably producing the best results. Controlling the gaze direction of the generated images, though, has proven more elusive. Part of the problem is that assessing gaze direction from an image, or at least determining whether it is congruent with that of another image, is difficult. Consider, for example, SimGAN [33], a popular algorithm that casts the synthesis problem as one of domain transfer. Starting from purely synthetic images, created using computer graphics from a model of the human eye, with prescribed gaze direction and head orientation, SimGAN generates realistic images sampled from a specific *target* domain. Gaze direction is controlled by adding to the adversarial loss a term that measures the L1 norm of the pixel-wise difference between the generated and the input image. Unfortunately, substantial photometric differences between the images in the two domains tend to bias this simple measure of gaze discrepancy, especially for larger image sizes. This is shown in the example of Fig. 1, wherein an image generated by SimGAN appears to look in a different direction than in the synthetic image provided in input.

Our approach to controlling the gaze direction of the generated images is inspired by the intuition that important information about gaze direction is revealed by a *segmen-*

tation mask of the eye image. A well-formed segmentation mask describes three main components of an eye image: the iris, the white sclera, and the skin area surrounding the sclera (see Fig. 1, lower row.) It is well known that, for a fixed head pose, the iris eccentricity (relative location of the iris within the white sclera) determines the perceived gaze direction [35], and that the amount of visible sclera depends on the head orientation [30]. It thus stands to reason that such a ternary mask could be used to represent gaze direction and head orientation. It is also conceivable that a well-designed segmenter should be able to extract acceptable segmentation masks from real eye images. Based on these observations, we decided to experiment with masks as *domain-independent proxies* for gaze direction.

Our proposed system takes in input a ternary mask produced by the UnityEyes graphic engine [39] with the desired head orientation and gaze direction, and generates an eye image with the same gaze direction in the "style" of the desired domain (Fig. 1, right column). The network is trained using a conditional GAN under the pix-to-pix paradigm [14]. Specifically, each training sample is formed by a pair (image, ternary mask) from the target domain. Whereas only the ternary mask is fed into the generator, the associated image is used for two purposes: to facilitate the job of the discriminator, and to enforce faithfulness of appearance by means of a L1 loss term that penalizes the difference between the input and the output images. Herein lies a critical difference with SimGAN: we never directly compare images from different domains, thus sidestepping the risk of bias from effects that are independent of gaze orientation.

A subtle but important characteristic of our algorithm is that the actual angles of gaze direction or head orientation are not needed at training time. We only use images from the target domain during training, and don't assume that these images have been annotated (as mentioned earlier, obtaining the required type of annotation can be challenging). Head pose and gaze direction information is embedded in the ternary masks, which are computed from the images themselves. When the generator is used to synthesize new images for a desired gaze direction, it takes in input a proper ternary mask, produced, for example, by UnityEyes.

For this system to work, it is critical that good quality ternary masks be available for images in the target domain. Standard segmentation algorithms can be used for this purpose, provided that enough labeled data is available for their training. Manual labeling (by drawing the iris and visible sclera regions in each image) is a conceivable option, albeit a time-consuming and error-prone one We decided instead to experiment with a training procedure that only uses the ternary masks automatically generated by UnityEyes along with the synthetic eye images. The segmenter is trained in parallel with the generator in an iterative fashion. This scheme is shown to produce excellent results after just a few iterations.

Our proposed EyeGAN system was evaluated comparatively in two different ways. First, we looked at the consistency of gaze direction by comparing the ternary masks computed on the generated images with the masks that were given as input. If the two masks agree, it can be expected that the perceived gaze direction of the generated images is congruent with the prescribed gaze direction, which was used to create the synthetic input. Second, we used EyeGAN to generate image data sets in target domains, while inheriting original annotations, and used this data to train networks for specific tasks: image region segmentation, pupil localization, and gaze direction estimation. These are applications of great interest for biometrics [27], medical diagnostic [36], and eye gaze tracking [28]. In many situations, annotating this type of data can be difficult or impossible, hence the interest in domain transfer methods for network training. The results of our experiments show that EyeGAN compares favorably with other state of the art domain transfer algorithms under the metrics considered.

2. Related Work

Due to its relevance in multiple application scenarios, the synthesis of realistic eye images has received considerable attention in the literature. Le *et al*. [25, 24] captured images under different head poses; eye images for new head poses were then synthesized via warping. Multiple cameras were used in [34] to build a 3D reconstruction of the eye region and to synthesize eye images for novel poses. Wood *et al*. [40, 39] rendered eye images (via computer graphics) using a 3D geometric eye model and head scans. This tool can be used to build very large data sets of perfectly annotated, high quality eye images. However, these synthetic images may not be representative of specific target domains, for which representative images may be available, but annotations may be difficult or impossible to obtain.

An approach to improving the quality of training data, while inheriting existing annotations, is to use a domain transfer algorithm. For example, SimGAN [33] transforms an eye image generated synthetically, with the desired head orientation and gaze direction, into a new image with the style of the target domain. This is accomplished by a GAN, trained to minimize an expected loss that includes two terms: the standard minimax adversarial loss (to ensure that the generated images look like samples from the target domain); and a L1 loss that penalizes discrepancies from the input synthetic eye image. This second term is meant to maintain consistency in gaze direction between the input synthetic image and the generated image. SimGAN produces impressive results, yet suffers from the problem that direct comparison of the generated and of the input image is difficult, as the images are from different domains.

Pixel-wise differences between the two images may thus be caused not only by a gaze direction discrepancy, but also by other irrelevant photometric factors (see e.g. Fig. 1).

In order to mitigate the problem of cross-domain comparison, Lee *et al*. [21] relied on the CycleGAN training procedure [43]. CycleGAN trains two generators, mapping images from source to the target domain, and vice-versa. A "cyclic loss" is defined (in additional to the standard adversarial loss) that penalizes the L1 norm of the difference between an image I in one domain, and the image obtained by mapping I to the other domain, then mapping the result back to the original domain. Hence, the L1 loss term is computed only between images in the same domain. Yet, this strategy alone cannot ensure that gaze direction is preserved. For example, the generator mapping images from the source to the target domain may still introduce a gaze direction discrepancy, provided that the generator from the target to the source domain learns to remove this discrepancy (that is, to "re-direct" gaze back to the original direction.) While CycleGAN maps source and target domains into separate latent spaces, other algorithms [20, 22] use a shared latent space for domain transfer from unpaired data. The method by Wang *et al*. [38] combine image synthesis and gaze estimation in a unified model.

Our EyeGAN system is directly inspired by the pix2pix algorithm for domain transfer [14]. Pix2pix requires pairs of images for training, where one image is from the source domain, and the other is the associated image in the target domain. A key insight of EyeGAN is that the generator does not need a highly detailed eye image input to produce a target domain image. What is needed is an input image with enough information to guide generation of a target domain image with the prescribed gaze direction. We use ternary mask images for this purpose. Closely related to our work is the Cycada algorithm [12], which used CycleGAN for domain transfer, then segmented the resulting images using a fully convolutional network (FCN) [23] (symultaneously trained), where the FCN loss is fed back into the GAN to ensure correctness of the inherited annotations. Differently from Cycada, our EyeGAN algorithm directly starts from a segmentation mask.

A related area of research is gaze redirection, wherein both input and output images are in the same domain. Initial work in this area aimed to learn a warp function to "turn" one's gaze to the desired direction [18, 8]. Improved results were recently obtained using adversarial training [11].

3. The EyeGAN Algorithm

Our system generates eye images with a desired style, as represented by a set of (un-annotated) images taken in a particular target domain. Head pose and gaze direction for the generated images are controlled by means of ternary masks which, as discussed in the Introduction, function as proxies for the desired pose and gaze direction. Specifically, we assume that a data set of synthetic images I_s (where the subscript $_s$ stands for "source") is available, along with associated masks M_s, also synthetically produced. (Alternatively, real images with manual mask annotations could be used.) At run time, the generator, implemented as a convolutional network, takes one such mask in input, and produces an image in the desired style. Note that, unlike similar algorithms such as SimGAN or CycleGAN, we do not use synthetically generated eye images as input, but only the associated masks.

The generator is trained according to two criteria: *(1) Realism:* the generated images must look realistic (as if they were actual samples from the target domain); *(2) Consistency:* the perceived gaze direction and head orientation of a generated image must conform to the prescribed values, in the sense that the associated mask should look similar to the mask fed into the generator. To generate realistic images, we follow the same conditional GAN strategy as pix2pix [14]. Specifically, the training data is formed by pairs mask–image in the target domain, of which only the mask is fed into the generator. The network is trained using a minimax adversarial loss, to which a L1 loss term is added to ensure that the generated image looks similar to the image associated with the input mask. This loss term enforces consistency: if the output image is similar (in L1 norm) to the image associated with the input mask, then the mask associated with the output image can be expected to be similar to the input mask. Critically, the L1 loss component is computed from two images that can be assumed to be from the same domain (unlike SimGAN). We used quadratic loss for the adversarial component, as it was shown to be superior to log loss in terms of training stability [26]. The overall loss function is thus:

$$\begin{aligned}\mathcal{L}(\mathcal{G},\mathcal{D}) = \mathbb{E}_{I_t}[D(I_t)-1]^2 + \mathbb{E}_{M(I_t)}[D(G(M(I_t)))]]^2 \\ +\lambda\,\mathbb{E}_{I_t}\|I_t - G(M(I_t))\|_1\end{aligned} \tag{1}$$

This training scheme requires availability of images I_t in the target domain along with associated masks M_t. Unfortunately, such masks are normally not available, and their production via manual labeling can be exceedingly time consuming. Instead, we create the required masks from target domain images using a properly trained semantic segmentation algorithm (such as FCN [23]) that takes in an image I_t to produce a mask $M(I_t)$ (note the overloaded use of the symbol M). Still, the problem remains: in order to train the segmenter, we need image–mask pairs. We tackle this problem by leveraging the pairs (I_s, M_s) available in the synthetic eye image data set. Intuitively, a segmenter trained on this data should be able to produce a recognizable, albeit probably not accurate, masks when applied to a target domain image. An example is shown in Fig. 3.

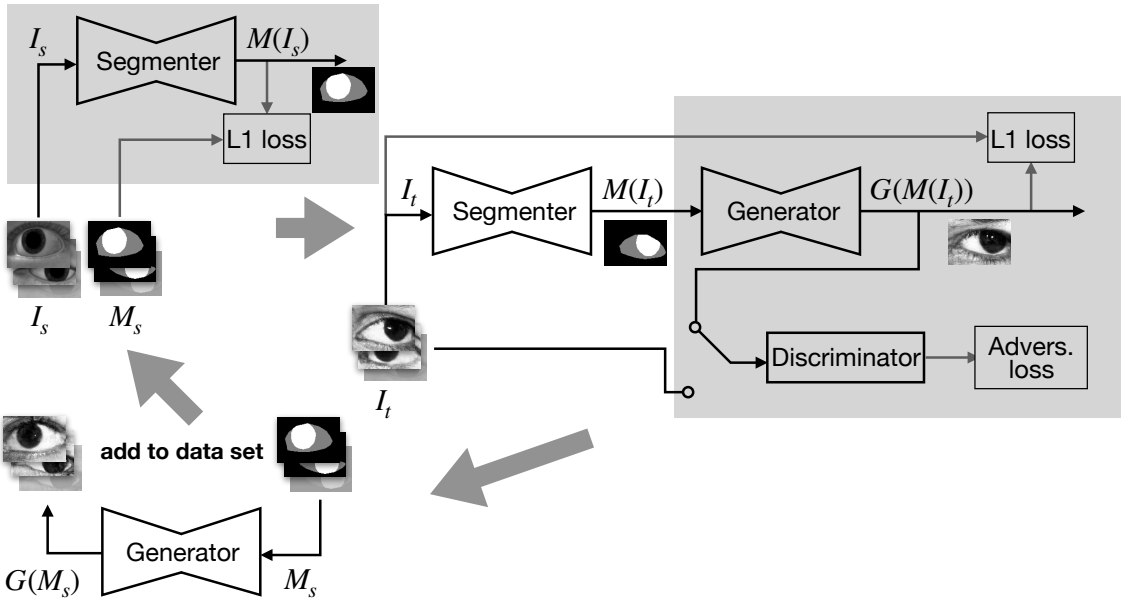

Figure 2. The overall training scheme of EyeGAN. At each step, the modules being trained are shown on a grey background.

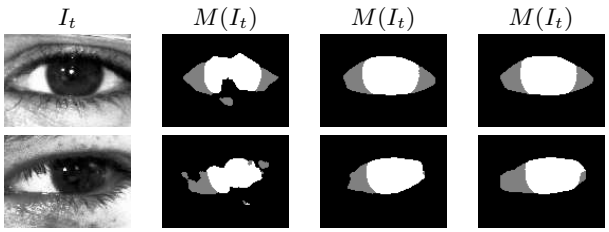

Figure 3. Two examples of segmentation of target domain images I_t. At Iteration 1, the segmenter was only trained using synthetic images and masks (I_s, M_s). In further iterations, pairs $(G(M_s), M_s)$ were added to the data set.

In order to improve the quality of segmentation, we augment the training data set for the segmenter (Fig. 2, top left) with image–mask pairs from the target domain. Of course, no improvement should be expected by simply adding to the training set pairs $(I_t, M(I_t))$, where the masks $M(I_t)$ were obtained using a suboptimal segmenter. Rather, we add pairs image-mask of the form $(G(M_s), M_s)$, where the masks M_s come from the synthetic data set (and thus are of perfect quality), and the associated images $G(M_s)$ are created by the generator, which, as explained earlier, was trained using pairs $(I_t, M(I_t))$. While not "real", these images can be considered to be samples from the distribution of the target domain (thanks to adversarial training.) We have observed that, after retraining the segmenter with this augmented data set, its performance on the target domain images improve noticeably (see Fig. 3). The process is then repeated. After 2–3 iterations, the segmenter produces satisfactory results, leading to good quality target domain image-mask pairs, which are used to re-train the generator. Fig. 2 shows the overall training scheme. Note that, at run time, the generator is only fed with synthetic mask M_s.

3.1. Implementation Details

In all of our experiments, source eye images and masks were created using the UnityEyes tool [39]. The images were cropped to only include the eye region, and resized to 120×88 pixels. The ternary masks were obtained from the landmark points provided to indicate the boundary the sclera and of the iris regions. A set of 25,000 synthetic images and masks was thus generated.

The segmenter was implemented using the FCN-8s architecture [23]. The learning rate was set to 0.001, with batch size of 8. The network was optimized using Adam [17]. A pytorch implementation[1] of the pix2pix scheme [14] was used to train the generator mapping masks to target domain images. The architecture of the generator was similar to that of [15, 43] with six Resnet [10] blocks. The discriminator used the same PatchGAN architecture of [43]. The balancing coefficient λ was set to 40.

4. Experiments

4.1. Gaze Direction Validation

A simple way to evaluate whether a generated eye image in the target domain (*target image* for short) is consistent with a desired gaze direction, is to compare its associated ternary mask (obtained via segmentation) with the mask fed

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

into the generator, which was synthetically created according to the prescribed gaze direction. If the two masks are identical, we may assume that gaze direction is maintained (more precisely, the gaze direction as perceived when observing the image coincides with the gaze direction represented by the input mask). A target image $I_t = G(M_s)$ whose mask $M(I_t)$ (as computed by the segmenter) is dissimilar from the input mask M_s, is unlikely to be judged to have the same gaze direction. We measure the similarity S of two equally-sized ternary masks M_1, M_2 by the number of pixels in which the masks agree, divided by the total number of pixels in each masks. The number $S(M_1, M_2)$ takes values between 0 and 1, and is equal to 1 only when the masks are identical. When considering the similarity of two masks, one synthetically produced[2] for gaze direction θ (denoted by M_s^θ), the other obtained by segmentation of the target image generated with input mask M_s^ϕ ($M(I_t^\phi)$, where $I_t^\phi \equiv G(M_s^\phi)$), we will use the shorthand $S_{s,t}(\theta, \phi) \equiv S(M_s^\theta, M(I_t^\phi))$.

We frame gaze orientation validation in probabilistic terms by defining a probability density function on the gaze direction θ perceived upon observation of a target image generated under prescribed gaze direction ϕ: $p(\theta|I_t^\phi)$. This means that, upon observing the target image I_t^ϕ, with probability $p(\theta|I_t^\phi)d\theta$ the perceived gaze direction angle is within an interval $d\theta$ around θ.

We will make the assumption that $p(\theta|I_t^\phi)$ is a function of the similarity between the mask $M(I_t^\phi)$, and the "ideal" mask for gaze direction θ, which is M_s^θ. Formally:

$$p(\theta|I_t^\phi) = K(\phi) f(S_{s,t}(\theta, \phi)) \tag{2}$$

where $f(S) = \exp(-\alpha \cdot (1 - S))$. α is a parameter that controls the dispersion of the density $p(\theta|I_t^\phi)$ (we set α=10 in our experiments.) $K(\phi)$ is a normalization constant that can be estimated as follows. We sample N gaze directions $\{\theta_i\}$ uniformly within the angular interval Θ in which θ can take values, and compute the mean $\bar{f}_\phi = \sum_{i=1}^N f(S_{s,t}(\theta_i, \phi))/N$:

$$\bar{f}_\phi \approx \frac{1}{K(\phi)} E_{\theta \sim \mathcal{U}(\Theta)}[p(\theta|I_t^\phi] = \frac{1}{K(\phi)\|\Theta\|} \tag{3}$$

from which we obtain:

$$K(\phi) = 1/(\bar{f}_\phi \cdot \|\Theta\|) \tag{4}$$

Given a target image generated for gaze angle ϕ, the probability that the perceived gaze direction coincides with ϕ with tolerance $d\phi$ is $p(\phi|I_t^\phi)d\phi$. Hence, the probability that the perceived gaze direction for a generic target image is "correct" (coinciding with the prescribed gaze direction, which is assumed to be uniformly distributed) within tolerance $d\phi$, is $p(C_{s,t})d\phi$, with:

$$p(C_{s,t}) = E_{\phi \sim \mathcal{U}(\Theta)}[p(\phi|I_t^\phi)] \tag{5}$$

$$\approx \frac{1}{N}\sum_{j=1}^N p(\phi_j|I_t(\phi_j)) = \frac{1}{N}\sum_{j=1}^N K(\phi_j) f(S_{s,t}(\phi_j, \phi_j))$$

$$\approx \frac{1}{\|\Theta\|}\sum_{j=1}^N \frac{f(S_{s,t}(\phi_j, \phi_j))}{\sum_{i=1}^N f(S_{s,t}(\theta_i, \phi_j))}$$

$$= \frac{1}{\|\Theta\|}\sum_{j=1}^N \frac{e^{-\alpha(1-S_{s,t}(\phi_j,\phi_j))}}{\sum_{i=1}^N e^{-\alpha(1-S_{s,t}(\theta_i,\phi_j))}}$$

where the prescribed gaze directions $\{\phi_j\}$ are sampled uniformly within Θ.

The relative effectiveness of different eye image synthesis methods at preserving gaze direction can be quantified by comparing $p(C_{s,t})$, computed for each method, with the same quantity computed in the "ideal" case, where $M(I_t)$ is substituted by M_s (the resulting value is denoted by $p(C_{s,s})$). The ratio $p(C_{s,t})/p(C_{s,s})$ (termed *gaze consistency index*) is shown for the SimGAN, CycleGAN, and EyeGAN methods in Tab. 1 (note that term $\|\Theta\|$ disappears in the ratio.) We used the segmenter designed as part of the EyeGAN training process to extract masks from the target images in all three methods. These results were obtained using Eq. (5) on $N = 81$ input masks M_s (or associated synthetic images I_s in the case of SimGAN and CycleGAN), sampled uniformly in terms of gaze direction and head orientation. Target domain images were culled from the UBIRIS data set [29], selecting those taken at a distance of 4 meters. The images were resized to 120×80 pixels and histogram equalized. The results show that EyeGAN produces a substantially higher gaze consistency index than the other methods.

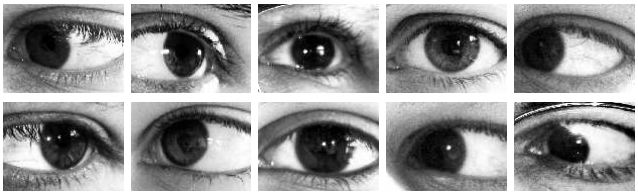

Figure 4. Sample images from the UBIRIS data set [29] (selected from those taken at a distance of 4 meters.) The images were histogram equalized.

$\mathbf{p(C_{s,t})/p(C_{s,s})}$		
EyeGAN	CycleGAN	SimGAN
0.89	0.36	0.48

Table 1. Gaze consistency indices for the methods considered.

[2]For simplicity of exposition, we only consider here one angle, instead of two, of gaze direction, and conflate head orientation with gaze direction.

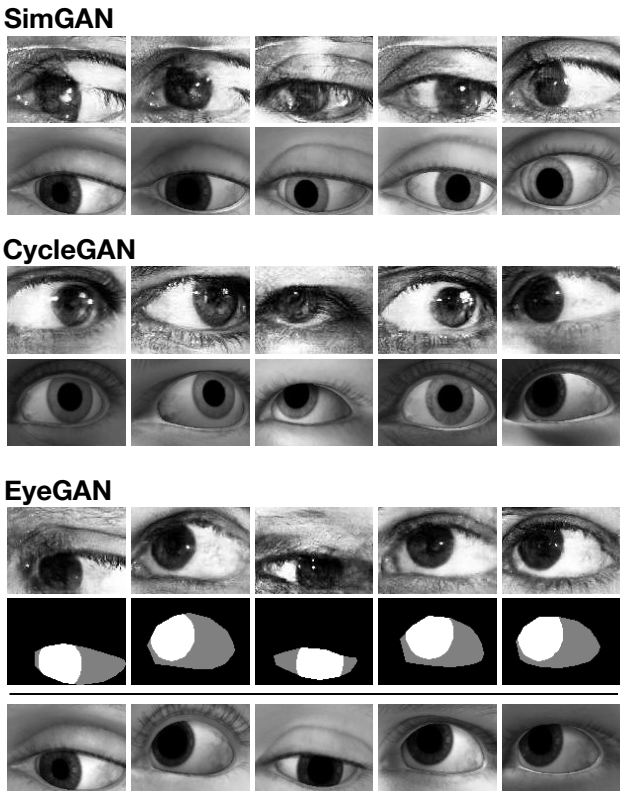

Figure 5. The five generated eye images with the lowest similarity score $S_{s,t}(\phi, \phi)$ for each method. Each image is shown with the synthetic image I_s or mask M_s that was fed into the corresponding generator. For reference, we also show the synthetic images I_s corresponding to the masks M_s for the EyeGAN case in the last row, even though only the masks M_s were fed into the generator in this case.

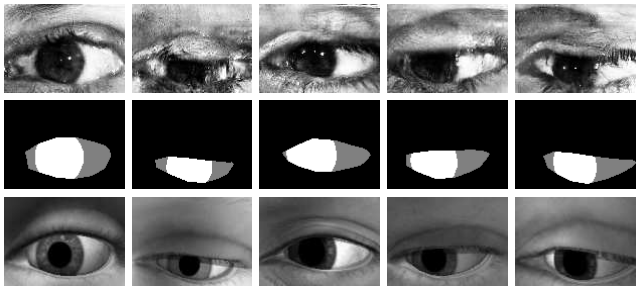

Figure 6. Examples of poor quality images generated by EyeGAN.

Examples of generated images for the three methods considered are shown in Fig. 5. For each method we selected the five images I_t with the lowest similarity score $S_{s,t}(\phi, \phi)$. Each image is shown next to the source image I_s or mask M_s (for EyeGAN) that was fed into the generator. We noted that EyeGAN generally produces images with better overall quality than the other two methods. Some examples of poor quality images generated by EyeGAN are shown in Fig. 6.

4.2. Eye Region Segmentation

Segmentation of various ocular regions as well as of the periocular region is instrumental for ocular biometric applications [27, 2]. Eye segmentation is also useful for animation of eyes and eyebrows of avatars for virtual reality [41]. Generation of training images via manual eye region segmentation and labeling, however, can be time consuming and thus expensive, and possibly error-prone. Domain transfer techniques can be used to generate large data sets with inherited annotations from labeled source domains. We comparatively evaluated our EyeGAN network as a tool to generate annotated training data for a segmenter tasked with extracting specific regions in eye images.

We considered two available labeled data sets for these experiment. The first data set (UBIRIS, already considered in Sec. 4.1) has manual annotations of the iris region. The second data set (SBVPI [31, 32]) contains 1822 eye images of 55 subjects looking towards four different directions (images were resized to 120×88 pixels.) SBVPI contains iris and pupil annotation for only a small number of subjects, but sclera and periocular masks are available for all subjects. Since both sclera and iris lie within the periocular region, the iris mask can be easily obtained as the area inside the periocular region that is not part of the sclera. Thus, for images in the SBVPI, we are able to access ground-truth ternary masks. Note that binary (for UBIRIS) and ternary (for SBVPI) masks were only used for validation (not during training).

Two subsets were culled from each data set, by partitioning the set of subjects associated with the images (i.e., any two eye pictures of the same subject were assigned to the same subset.) The first subset (1,750 images for UBIRIS, 1092 images for SBVPI) was used to train the domain–transfer network $G(M_s)$, while the remaining images in the considered data set were used to validate the segmenter, which was trained on images synthesized by EyeGAN using the synthetic masks.

Four different FCN segmenters were trained, where in all cases the labels were represented by synthetic masks M_s. Note that when experimenting with the UBIRIS data set, the ternary synthetic masks generated using UnityEyes were transformed into binary by conflating the sclera and background into one region. We first considered a baseline scenario, with the segmenter trained using the synthetic images I_s associated with the synthetic masks M_s, then tested on the real images. We then re-trained the segmenter using the same masks M_s as labels, but with domain–transferred images in input. These training images were generated starting from synthetic images ($G(I_s)$) using SimGAN and CycleGAN, and from synthetic masks ($G(M_s)$) for EyeGAN. All four segmenters were trained on 25,000 domain–transferred images. We used the metrics considered in [23] (Tab. 2 and Tab. 3) to evaluate the quality of segmentation. We also

show the results when the segmenter is trained directly on the real labels available in the training data ("train on target", or TT.) For both data sets, training the segmenter using target domain data produced by EyeGAN with inherited annotation from UnityEyes gave the best results. In fact, for the UBIRIS data set, the results using EyeGAN images for training are better than when the segmenter is trained on the real labels (TT). This can be justified by the fact that many more EyeGAN images (with inherited annotations) with variations in iris positions and size were available for training than real target images.

	Baseline	EyeGAN	CycleGAN	SimGAN	TT
IoU:Skin	0.95	**0.98**	0.93	0.96	0.97
IoU:Iris	0.77	**0.90**	0.68	0.80	0.87
mean IoU	0.86	**0.94**	0.81	0.88	0.92
f.w. IoU	0.92	**0.96**	0.89	0.93	0.96
pix. acc.	0.96	**0.98**	0.94	0.96	0.98
mean pix. acc.	0.90	**0.97**	0.88	0.93	0.94

Table 2. Comparison of segmentation into sclera and iris produced by the different algorithms considered for the UBIRIS data set, using standard metrics for multi-class segmentation [23], and specifically: IoU for each class; mean IoU; frequency weighted (f.w.) IoU; pixel accuracy; and mean pixel accuracy. The last column shows the "trained on target" (TT) results.

	Baseline	EyeGAN	CycleGAN	SimGAN	TT
IoU:Skin	0.86	**0.94**	0.84	0.83	0.96
IoU:Sclera	0.44	**0.78**	0.34	0.35	0.86
IoU:Iris	0.68	**0.84**	0.45	0.49	0.89
mean IoU	0.66	**0.85**	0.54	0.56	0.91
f.w. IoU	0.78	**0.91**	0.72	0.72	0.94
pix. acc.	0.87	**0.95**	0.82	0.82	0.97
mean pix. acc.	0.73	**0.92**	0.71	0.72	0.94

Table 3. Comparison of segmentation into skin, sclera, and iris produced by the different algorithms considered for the SBVPI data set. (See caption of Tab. 2.)

4.3. Pupil Localization

Another feature of interest in eye images is the location of the pupil center. High accuracy is needed for applications such as model–based gaze tracking [3]. We conducted an experiment similar to the one described in the previous section, where in this case the output of the network is a pair of numbers, representing the normalized coordinates of the estimated pupil center location. For this purpose, we used a DenseNet [13] architecture, with the last softmax layer replaced by a linear layer producing the coordinates vector. L2 loss was used for training. Specifically, we used the compact variant DenseNet-BC with the following configuration: L (number of layers) =100; k (growth rate of feature maps in each layer) =12; four dense blocks. The learning rate was set to 0.001 and the network parameters were optimized using Adam [17].

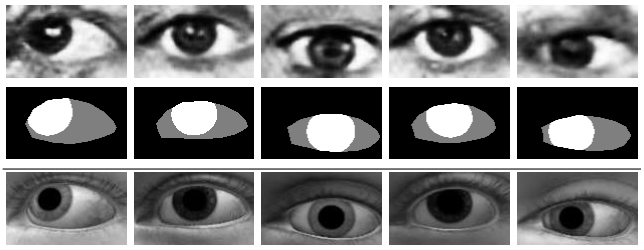

Figure 7. Examples of eye images generated by EyeGAN in the style of the BioID data set (top row), shown together with the UnityEye masks that were fed to the generator (middle row). For reference, we also show the synthetic images I_s corresponding to the masks M_s for the EyeGAN case in the last row.

As in the previous section, we trained a baseline regressor, using solely synthetic images generated by UnityEyes, as well as three regressors trained on domain–transferred images, inheriting annotations from UnityEyes. The target domain distribution was represented by the BioID data set [1], which contains 1521 grayscale images of 23 subjects taken at different head orientations. The images were resized to 120×72 pixels and histogram equalized. The location of the pupil center was available for each image; this information was only used in the final evaluation. Samples of the images produced by EyeGAN in the style of the BioID data set, generated starting from UnityEyes masks, are shown in Fig. 7. Fig. 8 shows the cumulative distribution function (CDF) of the Euclidean norm of the localization error for the different methods considered (where the CDF for a certain error value e represents the portion of images with error smaller than e.) For comparison, we also showed results using two well–known existing algorithms for pupil localization: ExCuSe [5] and ElSe [6], both based on fast elliptical fitting of the pupil region. Note that training with EyeGAN gave the best results.

4.4. Gaze Estimation

Appearance-based gaze estimation algorithms compute the direction of gaze directly from images of the user taken by a camera, without resorting to geometrical models of gaze formation. Training a network for appearance-based gaze estimation requires availability of data with precise annotation of gaze direction for each image. This can only be obtained indirectly, i.e. by asking the user to look at a certain point on the screen, and then inferring gaze direction from the known location of the user's head location. The ability to generate realistic images with inherited annotation is highly desirable, as it would enable construction of larger and more diverse training data sets.

We used NVGaze [16], a data set that contains both real

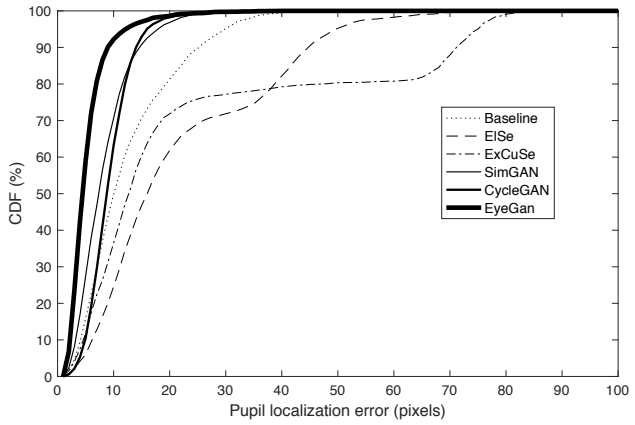

Figure 8. The cumulative distribution functions (CDFs) of the Euclidean norm of pupil localization error for the different algorithms considered (Sec. 4.3.)

eye images captured under IR illumination from a wearable headset device, as well as synthetic eye images from a similar viewpoint. The real eye images are annotated with gaze direction; the synthetic images have both gaze and segmentation mask annotation. The goal of this experiment was to learn a mapping from real images to gaze direction, but without making use of any available ground-truth labels for these images during training. We decided to use an intermediate representation, formed by a set of 25 feature points extracted from the segmentation masks (12 points uniformly distributed around the edge of the perioocular region, 12 points around the iris, and one point in the center of the iris.) We trained a fully connected neural network (with two hidden layers of size 500 each) to learn a mapping from feature points extracted from the synthetic masks to gaze direction. We then trained another network to predict the location of feature points from images generated with EyeGAN. During training, each EyeGAN image was associated with the feature points extracted from the synthetic mask used to generate the same image. This network used the same DenseNet [13] architecture described in Sec. 4.3, this time with a 50-dimensional (25×2) output (using L2 loss.) We then tested our system on real eye images using a cascade of the two networks just described: for each image, we first predicted the associated feature points, and then, from these feature points, the gaze direction.

The synthetic image portion of NVGaze contains two million images, while real eye images are collected for 35 subjects take at a high frame rate. For synthetic data, we sampled 50000 images. For real world data, we randomly selected one subject for with 76000 images containing 50 gaze directions, We sampled 128 images, ensuring that all gaze directions were covered. When testing the gaze detector on the real eye images, we reserved 22 such images for subject calibration [16]. This procedure, akin in spirit to subject calibration for standard IR gaze tracker, is designed to remove individual bias (as due, for example, to the kappa angle between the visual and the pupillary axes [19]. Specifically, we computed a quadratic regression from the predicted gaze directions to the ground-truth gaze directions over these 22 images; we then applied the same quadratic function on the predicted gaze for the remaining images, before computing the angular error with respect to the ground-truth gaze direction. The results are shown in Tab. 4, where "Baseline" represents the case in which the predictor for the feature points was trained entirely on synthetic data. This experiment once more shows that training the network (in this case, the feature points predictor) on EyeGAN-generated images with inherited annotations results in the best performance.

Baseline	EyeGAN	CycleGAN	SimGAN
23°	**5.3°**	20°	16°

Table 4. Mean gaze angular errors for the experiment described in Sec. 4.4.

5. Conclusions

We have introduced a new algorithm, EyeGAN, for the generation of eye images with a prescribed gaze direction in the style of a desired target domain. Like similar techniques, EyeGAN operates within the framework of domain transfer: starting from synthetically generated data, it produces an image that can be considered as a sample from the target domain distribution. The key difference between EyeGAN and other competing algorithms is in the way it enforces consistency of gaze direction. Our experiments have shown that ternary masks, which are easy to generate, contain enough information to "guide" the generation process into producing realistic images with the desired gaze direction. Comparative tests with different tasks and different target domains have shown that the images produced by EyeGAN lead to better results when used as training data with inherited annotations.

References

[1] The BioID database. www.bioid.com/facedb/. 7

[2] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn. Image understanding for iris biometrics: a survey. *Computer Vision and Image Understanding*, 110(2):281–307, 2008. 6

[3] J. Chen and Q. Ji. 3D gaze estimation with a single camera without ir illumination. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008. 1, 7

[4] A. Criminisi, J. Shotton, A. Blake, and P. H. Torr. Gaze manipulation for one-to-one teleconferencing. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 3, pages 13–16, 2003. 1

[5] W. Fuhl, T. Kübler, K. Sippel, W. Rosenstiel, and E. Kasneci. ExCuSe: Robust pupil detection in real-world scenarios. In *International Conference on Computer Analysis of Images and Patterns*, pages 39–51. Springer, 2015. 7

[6] W. Fuhl, T. C. Santini, T. Kübler, and E. Kasneci. ElSe: ellipse selection for robust pupil detection in real-world environments. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 123–130. ACM, 2016. 7

[7] K. A. Funes Mora, F. Monay, and J.-M. Odobez. EYEDIAP: a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 255–258. ACM, 2014. 1

[8] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European Conference on Computer Vision*, pages 311–326. Springer, 2016. 3

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[11] Z. He, A. Spurr, X. Zhang, and O. Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. *arXiv preprint arXiv:1903.12530*, 2019. 3

[12] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017. 3

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 7, 8

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 2, 3, 4

[15] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 4

[16] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 550. ACM, 2019. 7, 8

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 7

[18] D. Kononenko and V. Lempitsky. Learning to look up: Realtime monocular gaze correction using machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4667–4675, 2015. 3

[19] E. C. Lee and K. R. Park. A robust eye gaze tracking method based on a virtual eyeball model. *Machine Vision and Applications*, 20(5):319–337, 2009. 8

[20] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018. 3

[21] K. Lee, H. Kim, and C. Suh. Simulated+unsupervised learning with adaptive data generation and bidirectional mappings. In *International Conference on Learning Representations*, 2018. 3

[22] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 3

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 3, 4, 6, 7

[24] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Head pose-free appearance-based gaze sensing via eye image synthesis. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1008–1011. IEEE, 2012. 2

[25] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Gaze estimation from eye appearance: A head pose-free method via eye image synthesis. *IEEE Transactions on Image Processing*, 24(11):3680–3693, 2015. 2

[26] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017. 3

[27] I. Nigam, M. Vatsa, and R. Singh. Ocular biometrics: A survey of modalities and fusion approaches. *Information Fusion*, 26:1–35, 2015. 2, 6

[28] S. Park, X. Zhang, A. Bulling, and O. 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*, page 21. ACM, 2018. 2

[29] H. Proenca, S. Filipe, R. Santos, J. Oliveira, and L. A. Alexandre. The UBIRIS.v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1529–1535, 2009. 5

[30] P. Ricciardelli and J. Driver. Effects of head orientation on gaze perception: How positive congruency effects can be reversed. *The Quarterly Journal of Experimental Psychology*, 61(3):491–504, 2008. 2

[31] P. Rot, Ž. Emeršič, V. Struc, and P. Peer. Deep multi-class eye segmentation for ocular biometrics. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–8. IEEE, 2018. 6

[32] P. Rot, M. Vitek, K. Grm, Ž. Emeršič, P. Peer, and V. Štruc. *Handbook of Vascular Biometrics*, chapter Deep Sclera Segmentation and Recognition. Springer, 2019. 6

[33] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 2107–2116, 2017. 1, 2

[34] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, 2014. 2

[35] D. Todorović. Geometrical basis of perception of gaze direction. *Vision research*, 46(21):3549–3562, 2006. 2

[36] M. Tomasi, S. Pundlik, K. E. Houston, and G. Luo. Mobile device application for ocular misalignment measurement, Feb. 14 2019. US Patent App. 16/076,592. 2

[37] K. Wang and Q. Ji. Real time eye gaze tracking with 3D deformable eye-face model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1003–1011, 2017. 1

[38] K. Wang, R. Zhao, and Q. Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 440–448, 2018. 3

[39] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. 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. ACM, 2016. 1, 2, 4

[40] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3756–3764, 2015. 2

[41] Z. Wu, S. Rajendran, T. van As, J. Zimmermann, V. Badrinarayanan, and A. Rabinovich. Eyenet: A multi-task network for off-axis eye gaze estimation and user understanding. *arXiv preprint arXiv:1908.09060*, 2019. 6

[42] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2015. 1

[43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 3, 4