Geometry-Aware Eye Image-To-Image Translation

Conny Lu*
University of North Carolina at Chapel Hill
USA
connylu@cs.unc.edu

Qian Zhang*
University of North Carolina at Chapel Hill
USA
qzane@cs.unc.edu

Kapil Krishnakumar
Reality Labs, Meta
USA
kapilk@fb.com

Jixu Chen
Reality Labs, Meta
USA
jixu.chen@oculus.com

Henry Fuchs
University of North Carolina at Chapel Hill
USA
fuchs@cs.unc.edu

Sachin Talathi
Reality Labs Research, Meta
USA
stalathi@fb.com

Kun Liu
Reality Labs, Meta
USA
kun.liu@fb.com

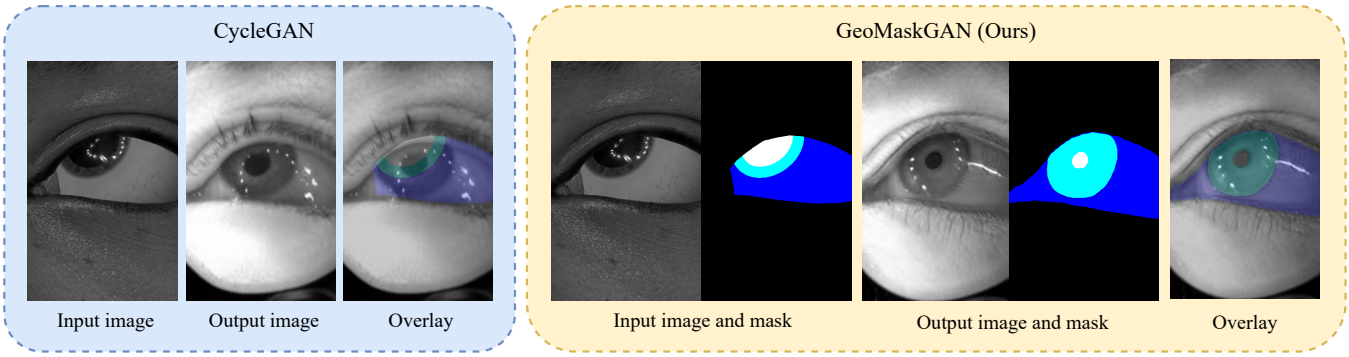

Figure 1: The original image-to-image translation method *(left)* generates an image with clear geometry change to close the geometric gap between two domains, but the generated image can no longer maintain the correspondence with the geometric-related ground truth (i.e. synthetic mask) as seen in the overlay. In contrast, our method utilizes an additional mask to generate both the high-fidelity image and aligned geometry represented as an output mask *(right)*.

ABSTRACT

Recently, image-to-image translation (I2I) has met with great success in computer vision, but few works have paid attention to the geometric changes that occur during translation. The geometric changes are necessary to reduce the geometric gap between domains at the cost of breaking correspondence between translated images and original ground truth. We propose a novel geometry-aware semi-supervised method to preserve this correspondence while still allowing geometric changes. The proposed method takes a synthetic image-mask pair as input and produces a corresponding real pair. We also utilize an objective function to ensure consistent geometric movement of the image and mask through the translation. Extensive experiments illustrate that our method yields a 11.23% higher mean Intersection-Over-Union than the current methods on the downstream eye segmentation task. The generated image has a 15.9% decrease in Frechet Inception Distance indicating higher image quality.

*Both authors contributed equally to this research.

ETRA '22, June 8–11, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9252-5/22/06…$15.00
https://doi.org/10.1145/3517031.3532524

CCS CONCEPTS

• **Computing methodologies** → **Image processing**; *Semi-supervised learning settings*; *Neural networks*; **Tracking**.

KEYWORDS

eye segmentation, eye tracking, syn2real, image-to-image translation, geometry consistency

ACM Reference Format:
Conny Lu, Qian Zhang, Kapil Krishnakumar, Jixu Chen, Henry Fuchs, Sachin Talathi, and Kun Liu. 2022. Geometry-Aware Eye Image-To-Image Translation. In *2022 Symposium on Eye Tracking Research and Applications (ETRA '22), June 8–11, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3517031.3532524

1 INTRODUCTION

Eye segmentation, distinguishing the main parts of the eye including the pupil, iris, and sclera, attracts considerable interest due to its fundamental role in eye tracking and raises an opportunity to investigate different types of eye movements. With the explosion of deep learning, the field of eye tracking has increasingly adopted deep learning based methods. However, due to its requirement for large amounts of training data, datasets collection has become a major challenge. This is particularly true in the scenario of Virtual/Augmented Reality, given the difficulty in equipment manufacturing [Ashtari et al. 2020]. The available datasets for VR/AR applications are collected using specific devices with fixed sensor configurations and the rapidly evolving design cycle for these devices limits the use of existing datasets. Data diversity is also not guaranteed due to financial, time, and geographical constraints. In addition, data annotation is time-consuming and labor-intensive, especially for pixel-level tasks (e.g. eye segmentation). One solution normally employed is to render synthetic datasets with powerful graphics engines. However, the domain gap between the synthetic and real data prevents the model trained on synthetic data to generalize well to real data.

To solve this problem, image-to-image translation [Fuhl et al. 2019; Huang et al. 2018; Isola et al. 2017; Lee et al. 2018; Liu et al. 2017; Zhu et al. 2017] can map from a synthetic domain to a real domain to enhance the fidelity of synthetic data. Unpaired image-to-image translation [Kim et al. 2017; Liu et al. 2017; Yi et al. 2017; Zhu et al. 2017] without requiring paired images as input simplifies the data acquisition process and is adopted in several applications [Alotaibi 2020; Pang et al. 2021]. However, one overlooked issue is that the object geometry in the image may change during the translation due to the natural distinction of geometric distribution in two domains. Taking the eye image as an example, the pupil size in the real domain may be statistically smaller than the one in the synthetic domain. The synthetic-to-real translation network then tends to match the smaller real pupil size. This problem is especially severe for geometric-related downstream tasks, such as eye segmentation and iris/pupil detection, since the geometric change pollutes the ground truth provided by the synthetic data, and thus reduces the task accuracy. Several recent works [Fu et al. 2019; Li et al. 2018; Xie et al. 2020] aimed to remedy this problem by proposing a soft gradient-sensitive objective for keeping the geometric boundaries during the translation [Li et al. 2018] or a content consistency loss punishing any geometry disparity by computing the L2 norm of two attention maps[Xie et al. 2020]. However, they strictly prevented the geometry of the image from changing during the translation, so the generated geometric distribution conforms to the synthetic domain, not the real domain. This gap that still remains in geometric distribution may lead to artifacts in the generated images and decreased accuracy when applying the generated datasets to subsequent tasks.

The motivation of our proposed work is to perform image-to-image translation such that the translated image data distribution is identical to the distribution of real data while maintaining the consistency between the generated image and mask. Specifically, in the eye segmentation task, we allow the eye geometry to change during the translation while preserving the geometric correspondence between the image and mask as shown in Figure 1. To this end, we propose GeoMaskGAN, a novel geometric-aware generative adversarial network [Goodfellow et al. 2014] based method that accepts a synthetic image and its corresponding semantic mask as input, and generates a pair of real image and mask. A triple-level adversarial loss is applied to preserve the fidelity of generated distribution, while a gradient loss is utilized to keep the correspondence between the generated image and its semantic label. We develop a semi-supervised algorithm using a limited amount of real masks to significantly boost the performance of our proposed I2I translation. Extensive qualitative and quantitative experiments demonstrate that our method achieves a 11.23% higher mIOU than the state-of-the-art in segmentation task and a 15.9% lower Frechet Inception Distance (FID) indicating better quality of generated images. We also conduct ablation studies to validate the effectiveness of different loss terms, discriminators, and the number of training masks.

2 METHOD

Our image-to-image translation method, GeoMaskGAN, has two main goals. First, the generated image distribution should be as close as possible to the real distribution, especially the distribution of geometric properties of the real data. Second, the geometric changes that happened to the translated images can be reflected in the translated semantic labels, which allows for using translated image-mask pairs as ground truth data for downstream tasks, such as semantic segmentation. To achieve these two goals at the same time, we generate an image with its underlying semantic label and ensure they both have a consistent geometric movement during the translation. Note, we do not require the generated image to perfectly preserve the geometry in the synthetic image like SG-GAN, which violates the first goal.

2.1 Problem formulation

Let $x^{\{R,S\}}$ represent a random variable drawn from the distribution $X^{\{R,S\}}$ of images in the real dataset and the synthetic dataset respectively. Let $g^{\{R,S\}}$ represent the hidden geometry of the image $x^{\{R,S\}}$ and its corresponding mask $m^{\{R,S\}}$. $g^{\{R,S\}}$ is a random variable drawn from the geometric distribution $G^{\{R,S\}}$ representing key eye geometric features such as the size and shape of the iris, pupil, and sclera.

The previous geometry-consistent image-to-image translation problem focused on solving for a functional mapping $f : x^S \sim X^S \rightarrow x^R \sim X^R$, without considering translating the geometric properties of the image, i.e., $g^R \sim G^S, g^R \nsim G^R$. In contrast, we are interested in solving the image-to-image translation problem where the image from the synthetic domain that is translated into the real domain can be constrained to match the geometric distribution in the real domain. In other words, we aim to solve a functional mapping: $f : x^S \sim X^S \rightarrow x^R \sim X^R; g^R \sim G^R$.

In contrast to unsupervised learning, we deploy a semi-supervised learning method to boost the performance using only a small amount of mask annotations in the real domain. In this case, we have a synthetic dataset $X^S = \{x_i^S, m_i^S\}_{i=1}^{n^S}$ with n^S synthetic image-mask pairs and a real dataset with a small amount of image-mask pairs $X^R = \{x_i^R, m_i^R\}_{i=1}^{n^R}$ but a large amount of unlabeled images $X^U = \{x_i^U\}_{i=1}^{n^U}$ ($n^U \gg n^R$).

2.2 Network architecture

Figure 2 illustrates the network architecture of our method based on CycleGAN. Different than CycleGAN, our input and output of generators are image-mask pairs instead of images. We also design a triple-level discriminator to take advantage of both labeled and unlabeled data, D_I in the image level, D_P in the image-mask pair level, and D_{Sem} in the semantic level. D_I encourages generated images to be more consistent with the distribution of real images by distinguishing between a real image x^R sampled from both X^R and X^U and a generated fake image $\hat{x}^R$. D_P implicitly considers more intra-pair information by distinguishing between a real pair $\{x^R, m^R\}$ sampled from X^R and a translated pair $\{\hat{x}^R, \hat{m}^R\} = T^{S\to R}(x^S, m^S)$. The semantic discriminator D_{Sem} is inspired by the semantic discriminator proposed in SG-GAN, which focuses on semantic specific information and reduces the interference between classes during the translation. Our results illustrate that the usage of additional unlabeled data and our triple-level discriminator can boost the generated image quality and the eye segmentation performance.

2.3 Loss functions

To synthesize image/semantic mask distribution with high fidelity while tracking the geometric change, we introduce an objective that consists of four loss terms: *a gradient loss*, which encourages the geometric match between generated images and generated masks; *an adversarial loss*, which aims to align the image distribution between the real and synthetic domain; *a style loss*, which helps generator to output segmentation mask with less noise, and *a cycle consistency loss*, which further reduces the space of possible mapping functions by making the reconstructed images the same as the original input images. To simplify, we only express loss functions applied to the real domain below, but in the training phase, loss functions are used in both synthetic and real domains.

Grad loss. Inspired by [Li et al. 2018], we consider preserving the geometric similarity between the generated images and masks by computing their gradients. We compute gradients by convolving images with derivative kernels only on mask semantic boundaries. We express the objective as:

$$L_{grad}^R = \mathbb{E}[||(abs(S * \hat{x}^R) - abs(S * \hat{m}^R)) \odot sgn(S * \hat{m}^R)||_1] \quad (1)$$

where $\{\hat{x}^R, \hat{m}^R\} = T^{S\to R}(x^S, m^S)$, sgn is the sign function, S is a derivative filter, $*$ is convolution operation. In the experiments, we choose Sobel filters [Kanopoulos et al. 1988] for generating sharper boundaries. Different from [Li et al. 2018], we represent the semantic mask as a grayscale image and carefully assign the color for each class, reducing the impact of the difference in color intensity on image and mask.

Adversarial loss. Based on the triple-level discriminator we introduced in Section 2.2, we apply adversarial losses [Goodfellow et al. 2014] to all three discriminators. The adversarial loss applied to the real domain is,

$$\begin{aligned} L_{adv}^R = {} & \mathbb{E}(log(D_P^R(x^R, m^R))) + \mathbb{E}(log(1 - D_P^R(\hat{x}^R, \hat{m}^R))) \\ & + \mathbb{E}(log(D_I^R(x^R))) + \mathbb{E}(log(1 - D_I^R(\hat{x}^R)) \\ & + \mathbb{E}(log(D_{Sem}^R(x^R, m^R))) + \mathbb{E}(log(1 - D_{Sem}^R(\hat{x}^R, \hat{m}^R))) \end{aligned} \quad (2)$$

Cycle consistency loss. We also apply the cycle consistency loss proposed in CycleGAN to stabilize the training based on the assumption that if we translate images from one domain to the other and back again we should arrive at the image in the first domain. The cycle consistency loss is,

$$L_{cyc}^R = \mathbb{E}[||T^{R\to S}(T^{S\to R}(x^S, m^S)) - (x^S, m^S)||_1] \quad (3)$$

It is worth mentioning that although the cycle consistency loss can roughly preserve the global geometric information during the translation as mentioned in [Zhu et al. 2017], in our observation, the detailed local geometry is hard to maintain, especially in datasets without huge variance but with significantly different geometric distributions in two domains. In our case, the cycle consistency loss preserves the approximate position of the eyes in the image, but the pupil size and eyelid shape can still move easily. Experiments show that combining the cycle consistency loss with other losses can achieve a much better result.

Style loss. One issue we observed in our experiments is that the generated masks sometimes contain a lot of square-shaped texture artifacts. To ameliorate this problem, we deploy a style loss commonly used in style transfer [Gatys et al. 2015] to ensure that the basic texture of input synthetic masks and the generated masks are similar. For specific details please refer to [Gatys et al. 2015].

Full Objective. The total loss function could be represented in equation 4, each loss term $L = L^R + L^S$.

$$L = \lambda_{grad} L_{grad} + \lambda_{cyc} L_{cyc} + \lambda_{adv} L_{adv} + \lambda_{style} L_{style} \quad (4)$$

where λ_{grad}, λ_{cyc}, λ_{adv}, and λ_{style} are the hyper-parameters used to control the relative importance of the loss terms. The final targets can be represented as:

$$T^* = \arg\min_T \max_D L \quad (5)$$

where $T = \{T^{S\to R}, T^{R\to S}\}$, $D = \{D_I^R, D_I^S, D_P^R, D_P^S, D_{Sem}^R, D_{Sem}^S\}$.

3 EXPERIMENTS

We validate the effectiveness of our GeoMaskGAN by training an I2I network to translate the synthetic eye data in RIT-Eyes [Nair et al. 2020] to the real eye data in OpenEDS2019 [Garbin et al. 2019]. We apply the translated results to an eye segmentation task to further validate the geometry-awareness ability of our method. We compare our method against several state-of-the-art works both qualitatively and quantitatively as well as perform ablation studies on the loss objectives, triple-level discriminator, and the number of ground truth masks used in the semi-supervised training.

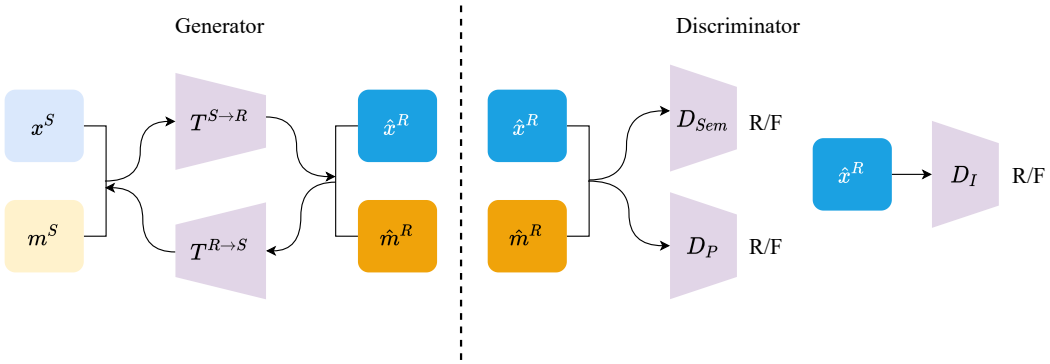

Figure 2: The network architecture of our method. First, we input a synthetic image x^S with its corresponding mask m^S into the generator $T^{S\to R}$ and generate a realistic pair $\{\hat{x}^R, \hat{m}^R\}$, which is then fed into an inverse generator $T^{R\to S}$ to obtain a reconstructed pair. We utilize a triple-level discriminator to distinguish the real (R) and generated (F) data from image level (D_I), image-mask pair level (D_P), and semantic level (D_{Sem}). Note the figure only describes the network of the forward loop from synthetic to real, the backward loop from real to synthetic is similar.

Table 1: Eye segmentation performance for different methods evaluated on OpenEDS2019 test set.

Method	Image FID ↓	Mask FID ↓
CycleGAN	101.9	-
SG-GAN	162.3	-
GcGAN	111.4	-
GeoMaskGAN-200 (Ours)	**86.0**	**122.0**
GeoMaskGAN-100 (Ours)	103.2	123.8
GeoMaskGAN-50 (Ours)	94.3	143.7

3.1 Experiments settings

Datasets. We choose OpenEDS2019 as our real dataset and S-openeds in RIT-Eyes as our synthetic dataset. OpenEDS2019 is an eye segmentation dataset of 12759 images with annotations at a resolution of 640×400, 8916 of them for training, 2403 for validation, and 1440 for testing. RIT-Eyes is a synthetic eye image generation platform, S-openeds in RIT-Eyes simulates the hardware configuration of OpenEDS, where 20 subjects with 41324 images are used for training and the remaining 4 subjects with 10330 images for validation. Both datasets contain four classes: pupil, iris, sclera, and background.

Evaluation metrics. We evaluate our method in terms of generated image and mask realism, and its utility to improve the performance of the eye segmentation task. We adopt the perception-based criterion Frechet Inception Distance (FID) [Heusel et al. 2017] to evaluate the image and mask quality. In addition, we train an eye segmentation network using the images generated by GeoMaskGAN and compute the segmentation accuracy on the test set of OpenEDS2019 including class Intersection-Over-Union (classIOU), mean Intersection-Over-Union (mIOU), and mean pixel accuracy.

Training details. We train our neural network on PyTorch [Paszke et al. 2017]. Similar to CycleGAN, we adopt resnet50 [He et al. 2016] and PatchGAN [Isola et al. 2017] as our generator and discriminator.

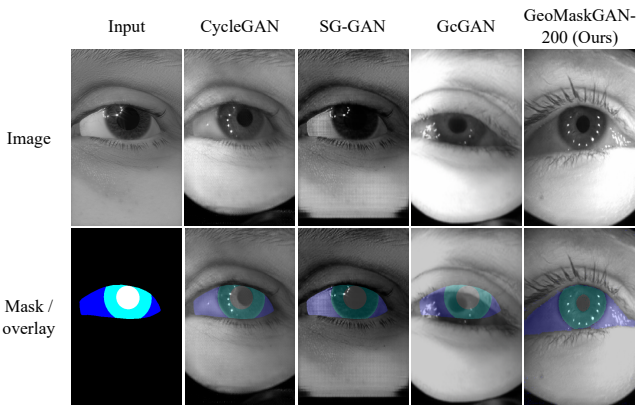

Figure 3: Comparison of qualitative results on image translation. The first column shows the input image and mask, other columns represent output images using different methods. The image-mask correspondence is visualized using the overlay of masks on images.

Our eye segmentation network is based on RIT-Net [Chaudhary et al. 2019]. For both generator and discriminator, we use the Adam solver [Kingma and Ba 2014] with a learning rate of 0.0002 and betas of (0.5, 0.999). We first grayscale the input images and then resize them to 256×256. The loss weights are set to $\lambda_{grad} = \lambda_{style} = 0.1$, $\lambda_{cyc} = \lambda_{adv} = 1.0$.

3.2 Comparison of state-of-the-art

Qualitative evaluation. We qualitatively compare our method with the state-of-the-art work CycleGAN [Zhu et al. 2017], SG-GAN [Li et al. 2018], GcGAN [Fu et al. 2019] to better demonstrate the superiority of our method. As shown in Figure 3, we first visualize the eye image-to-image translation results. The first column shows the input synthetic pair, while the other columns show the images generated by different methods (top row) and the overlay

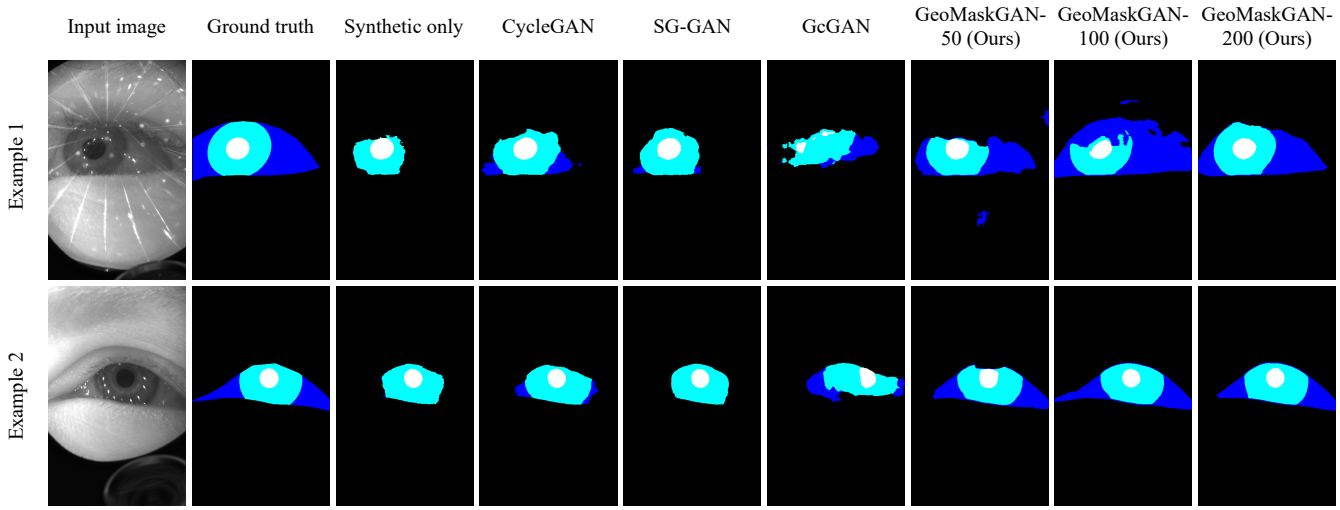

Figure 4: Qualitative results of eye segmentation. Each row represents one example of eye segmentation. The first and second column describes input images and ground truth segmentation masks, while other columns represent the segmentation results trained on data generated by different image generation methods.

Table 2: Eye segmentation performance for different methods evaluated on OpenEDS2019 test set.

Method	Bg IOU	Pupil IOU	Iris IOU	Sclera IOU	mIOU ↑	Pixel Accuracy ↑
Synthetic only	0.9110	0.8788	0.7385	0.2083	0.6838	0.7379
CycleGAN	0.9482	0.8957	0.8243	0.4768	0.7860	0.8268
SG-GAN	0.9219	0.8049	0.7663	0.2120	0.6760	0.7459
GcGAN	0.9106	0.1404	0.3498	0.2585	0.4150	0.5010
GeoMaskGAN-50 (ours)	0.9742	0.8070	0.8085	0.7050	0.8234	0.8708
GeoMaskGAN-100 (ours)	0.9694	0.8434	0.8273	0.7233	0.8406	0.9131
GeoMaskGAN-200 (ours)	**0.9845**	**0.9038**	**0.9041**	**0.8017**	**0.8983**	**0.9357**

of masks on images to describe the geometric consistency (bottom row). As CycleGAN, SG-GAN, and GcGAN output images only, we use the synthetic mask to compute the overlay. Our method can generate aligned image-mask pairs without dropping the image quality, while the images generated by CycleGAN and GcGAN have an untracked geometric movement. Although SG-GAN produces images in alignment with the synthetic masks, the geometry distribution of generated eye images deviates from the real eye geometric distribution. To further evaluate the effectiveness of the method on downstream tasks, we train a top-tier eye segmentation network RIT-Net [Chaudhary et al. 2019]. For our GeoMaskGAN, we train RIT-Net using the generated image-mask pairs. For methods that only generate images, we train RIT-Net using the generated images and original synthetic masks. The segmentation results on the OpenEDS2019 test set are shown in Figure 4, each row representing one example. Our method generates segmentation masks with higher accuracy and cleaner boundary, outperforming the state-of-the-art works, especially in sclera segmentation.

Quantitative evaluation. Table 1 describes the quality of images and masks generated by different methods in terms of FID. Our GeoMaskGAN using only 200 ground truth masks (GeoMaskGAN-200) achieves a 15.6% lower image FID than the state-of-the-art methods, indicating the higher image fidelity. Table 2 demonstrates the eye segmentation performance of different methods evaluated on the OpenEDS2019 test set. Our method also yields a 11.23% and 10.89% increase on mIOU and pixel accuracy, separately. We also observe that the sclera IOU is largely improved by our method, which indicates a big gap between the synthetic and real sclera. This finding casts a new light on the potential of our method to be used in guiding the generation of synthetic datasets. Another finding is that GcGAN even has a worse performance than training using only synthetic data (synthetic only), which shows that within a certain range the consistency between images and masks plays a more important role than image fidelity.

3.3 Ablation studies

We first conduct ablation studies on the gradient loss. In Table 3, we observe a significant decrease of mask FID from 155.2 to 122.0 after adding the grad loss, which indicates its ability to improve the mask fidelity, which can also be seen in the second and third column of Figure 5. The segmentation results with gradient loss

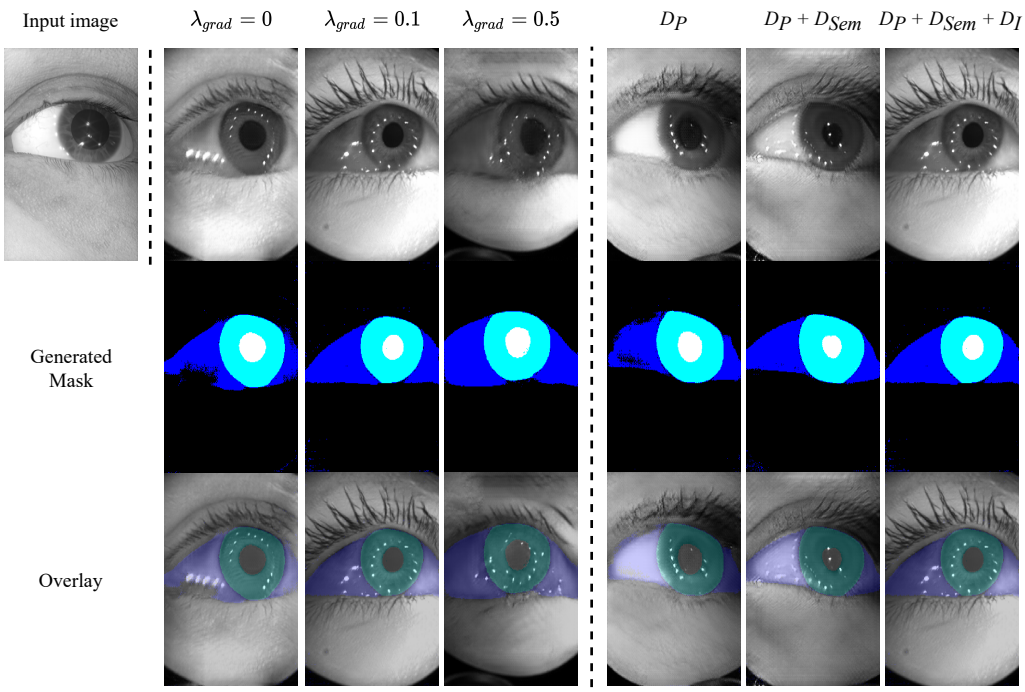

Figure 5: Qualitative results of ablation studies on different weights of the gradient loss and different combinations of our triple-level discriminator.

Table 3: Ablation studies on the grad loss term, evaluated by the image quality and segmentation accuracy.

Condition	Image FID ↓	Mask FID ↓	mIOU ↑
$\lambda_{grad} = 0$	**84.8**	155.2	0.8958
$\lambda_{grad} = 0.1$	86.0	**122.0**	**0.8983**
$\lambda_{grad} = 0.5$	98.5	124.2	0.8614

Table 4: Ablation studies on the triple-level discriminator, evaluated by the image quality and segmentation accuracy.

Condition	Image FID ↓	Mask FID ↓	mIOU ↑
D_P	114.4	130.5	0.7938
$D_P + D_{Sem}$	108.5	**90.9**	0.8483
$D_P + D_{Sem} + D_I$	**86.0**	122.0	**0.8983**

also produce a slightly better mIOU. We speculate this insignificant boost is because the pair discriminator D_P also provides geometric consistency capability. Moreover, setting a large weight of grad loss may result in a decrease in image quality. As shown in Figure 5, the boundary of images with $\lambda_{grad} = 0.5$ becomes more blurry than $\lambda_{grad} = 0.1$.

Table 4 also describes the results of using different discriminator combinations. After adding the semantic discriminator D_{Sem}, all the image FID, mask FID, and mIOU have a significant improvement which provides evidence that D_{Sem} can help in generating higher fidelity images by focusing more on the semantic information rather than the global texture. The image discriminator D_I is able to improve the image quality with additional unlabeled images, achieving a 32.3% lower image FID. The best performance is produced using our complete triple-level discriminator.

We also conduct experiments on evaluating the results of training with a different number of real masks (50, 100, 200). Table 1 and 2 show that only 50 labeled images can already achieve a higher mIOU and lower FID than current works. More labeled images keep improving the image quality and the segmentation results as shown in Figure 4 in terms of less noise on masks and cleaner boundaries.

4 CONCLUSION

We presented a semi-supervised eye image-to-image translation method that simultaneously generates images and masks with high fidelity while maintaining the geometric consistency between them. We introduced a neural network architecture that takes the pair of synthetic images and masks as input and generates a corresponding realistic image-mask pair. We improved the geometric correspondence between the generated image and mask using a gradient loss and a triple-level discriminator. The generated pair datasets allowed us to produce accurately labeled data as ground truth for downstream tasks such as eye segmentation. The qualitative and quantitative experiments showed that our method outperformed the state-of-the-art approaches in terms of both image-mask correspondence and image quality. Although our method can achieve compelling results, it still suffers from certain weaknesses. For example, some artifacts can be seen in the generated images in extreme

lighting conditions and with small eye-openings. Future research will explore more techniques used in semi-supervised learning, such as self-training and entropy minimization as possible solutions to resolve these artifacts. We hope our work can shed light on the importance of geometric movement during image-to-image translation and its impact on geometry-related downstream tasks, especially eye segmentation.

REFERENCES

Aziz Alotaibi. 2020. Deep generative adversarial networks for image-to-image translation: A review. *Symmetry* 12, 10 (2020), 1705.

Narges Ashtari, Andrea Bunt, Joanna McGrenere, Michael Nebeling, and Parmit K Chilana. 2020. Creating augmented and virtual reality applications: Current practices, challenges, and opportunities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

Aayush K Chaudhary, Rakshit Kothari, Manoj Acharya, Shusil Dangi, Nitinraj Nair, Reynold Bailey, Christopher Kanan, Gabriel Diaz, and Jeff B Pelz. 2019. RITnet: Real-time semantic segmentation of the eye for gaze tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 3698–3702.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. 2019. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2427–2436.

Wolfgang Fuhl, David Geisler, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2019. The applicability of Cycle GANs for pupil and eyelid segmentation, data generation and image refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

Stephan J Garbin, Yiru Shen, Immo Schuetz, Robert Cavin, Gregory Hughes, and Sachin S Talathi. 2019. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702* (2019).

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

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).

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*. 172–189.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. 1988. Design of an image edge detection filter using the Sobel operator. *IEEE Journal of solid-state circuits* 23, 2 (1988), 358–367.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*. PMLR, 1857–1865.

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

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*. 35–51.

Peilun Li, Xiaodan Liang, Daoyuan Jia, and Eric P Xing. 2018. Semantic-aware grad-gan for virtual-to-real urban scene adaption. *arXiv preprint arXiv:1801.01726* (2018).

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*. 700–708.

Nitinraj Nair, Aayush Kumar Chaudhary, Rakshit Sunil Kothari, Gabriel Jacob Diaz, Jeff B Pelz, and Reynold Bailey. 2020. Rit-eyes: realistically rendered eye images for eye-tracking applications. In *ACM Symposium on Eye Tracking Research and Applications*. 1–3.

Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. 2021. Image-to-Image Translation: Methods and Applications. *arXiv preprint arXiv:2101.08629* (2021).

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

Xinpeng Xie, Jiawei Chen, Yuexiang Li, Linlin Shen, Kai Ma, and Yefeng Zheng. 2020. Self-Supervised CycleGAN for Object-Preserving Image-to-Image Domain Adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 498–513.

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*. 2849–2857.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.