Proceedings of Machine Learning Research 95:863-878, 2018 ACML 2018

Refining Synthetic Images with Semantic Layouts by Adversarial Training

Tongtong Zhao ZHAOTONGTONG@DLMU.EDU.CN
Dalian Maritime University
Dalian 116026, China

Yuxiao Yan YUXIAOYAN@DLMU.EDU.CN
Dalian Maritime University
Dalian 116026, China

JinJia Peng PENGJINJIA@DLMU.EDU.CN
Dalian Maritime University
Dalian 116026, China

HaoHui Wei WEIHAOHUI@DLMU.EDU.CN
Dalian Maritime University
Dalian 116026, China

Xianping Fu FXP@DLMU.EDU.CN
Dalian Maritime University
Dalian 116026, China

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

Recently, progress in learning-by-synthesis has proposed training models on synthetic images, which can effectively reduce the cost of manpower and material resources. However, learning from synthetic images still fails to achieve the desired performance compared to naturalistic images due to the different distribution of synthetic images. In an attempt to address this issue, previous methods were to improve the realism of synthetic images by learning a model. However, the disadvantage of the method is that the distortion has not been improved and the authenticity level is unstable. To solve this problem, we put forward a new structure to improve synthetic images, via the reference to the idea of style transformation, through which we can efficiently reduce the distortion of pictures and minimize the need of real data annotation. We estimate that this enables generation of highly realistic images, which we demonstrate both qualitatively and with a user study. We quantitatively evaluate the generated images by training models for gaze estimation. We show a significant improvement over using synthetic images, and achieve state-of-the-art results on various datasets including MPIIGaze dataset.

Keywords:
Generative Adversarial Networks(GANs), Style Transfer, Learning-by-synthesis

1. Introduction

There is no denying that appearance-based gaze estimation has been enjoying its unique and significant role in facial information gathering via mass of labeled training datasets, which are gaining importance with the recent rise in high capacity deep convolution networks. However, due to the high cost of time and bankroll, solutions are required to tackle these problems. When it comes to this matter, human give priority to the synthetic image because the annotations are automatically available. However, due to the gap between synthetic and naturalistic image distributions, learning the misleading synthetic images can result in synthetic data not being a true reflection of realism, and the details represented may confuse the network and render it fail to complete the mission.

As such, one solution is to improve the simulator. But increasing the authenticity is computationally expensive, designing a renderer is a heavy workload, and the top renderer may still be difficult to model all the features of the naturalistic image. This may make the model over fitting in the "unreal" details of the synthetic image. The other solution is to improve the distribution of synthetic images and make them closer to the real pictures. The current method of state-of-the-art is Shrivastava et al. (2016). We adopt a neural network model similar to Generative Adversarial Networks (GAN). The main use of GAN was to train computers to generate some emanational pictures. To be graphic, it used a synthetic-image-producing network to be against another dataset that produced naturalistic images, and then distinguished it with a separate distinction network. On the basis of GAN, they made some big difference on models. For example, they input synthetic images instead of random vectors and proposed a learning model called Simulated + Unsupervised ultimately.

The contribution of this paper to computer vision, in addition to a new learning model, also includes using the model successfully train an optimized network (Refiner) on the premise of no artificial annotation and rendering computers generate more real synthetic images. However, the disadvantage of the method is that the distortion is not improved and the authenticity level is not stable. So, to solve this problem, we put forward a new structure, which can improve synthetic images, via the reference to the idea of style transformation to efficiently reduce the distortion of pictures and minimize the need of real data annotation. The reason why real data needs a small part of the annotation is that it needs its semantic information to make the synthetic data more authentic, while one of the great benefits of synthetic data is that its semantic information is clearer. For example, data sets such as unity of human eyes can use existing information to achieve accurate segmentation of pupil and iris. The advantage of applying this segmentation result to simulation data synthesis is that the addition of semantic information will make the distributed learning more apposite compared to holistic image synthesis and, in result, avoid the edge and pattern distortion caused by holistic learning. The same as general GAN structure, our framework also includes the generation network G and the distinction network D. We improve the structure of the image generation part and change the input from the random vector to the content of naturalistic image distribution and the simulation picture together. It will make the generation more stable, avoiding the randomness of distribution. It will also achieve a stable distribution in a short time. We modify the way of loss evaluating of the distinction network and add regular items to ensure the authenticity of the pictures.

We prove that the structure can generate highly realistic images steadily by qualitative and user research. Meanwhile, the training model of gaze estimation is used to evaluate produced images quantificationally. Compared with the synthetic images used, we implemented the best results on multiple datasets.

In summary, our contributions are five-fold:

1. We propose a new structure, which can improve synthetic images, via the reference to the idea of style transformation to efficiently reduce the distortion of pictures and minimize the need of real data annotation.

2. One of the great benefits of synthetic data is that its semantic information is clearer. The advantage of applying this segmentation result to simulation data synthesis is that the addition of semantic information will make the distributed learning more apposite compared to holistic image synthesis and, in result, avoid the edge and pattern distortion caused by holistic learning.

3. We improve the structure of the image generation part and change the input from the random vector to the content of naturalistic image distribution and the simulation picture together. It will make the generation more stable, avoiding the randomness of distribution. It will also achieve a stable distribution in a short time.

4. We modify the way of loss evaluating of the distinction network and add regular items to ensure the authenticity of the pictures.

5. We prove that the structure can generate highly realistic images steadily by qualitative and user research. Meanwhile, the training model of gaze estimation is used to evaluate produced images quantificationally. Compared with the synthetic images used, we implemented the best results on multiple datasets.

2. Proposed Method

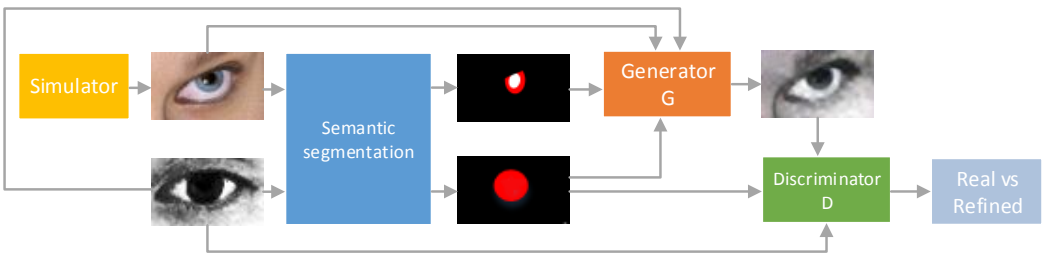

Figure 1: The overview of proposed methods. The proposed network can be divided into three parts: coarse segmentation network, Generator and Discriminator.

Our proposed network (As Fig. 1) takes two images with their mask: the reference style image which is a set of naturalistic eye image from video of driving environment or naturalistic eye image dataset. A stylized and retouched image referred as the input image from synthetic image dataset. We use this to train the gaze estimation, as we seek to transfer the style of the reference to the input while keeping the content and spatial information

due to its importance in appearance-based gaze estimation. The proposed network can be divided into three parts: coarse segmentation network, Generator and Discriminator.

2.1. semantic segmentation

We train the semantic segmentation network which builds upon an efficient redesign of convolutional blocks with residual connections to segment, according to the line of gaze estimation for the naturalistic image. One of the great benefits of synthetic data is that its semantic information is clearer. Thus the challenge is mainly on segment naturalistic image. Residual connections can avoid the degradation problem with a large amount of stacked layers. Our architecture is fully depicted in fig.2. *Number of feature maps at layers @ output resolution* is shown under each block.

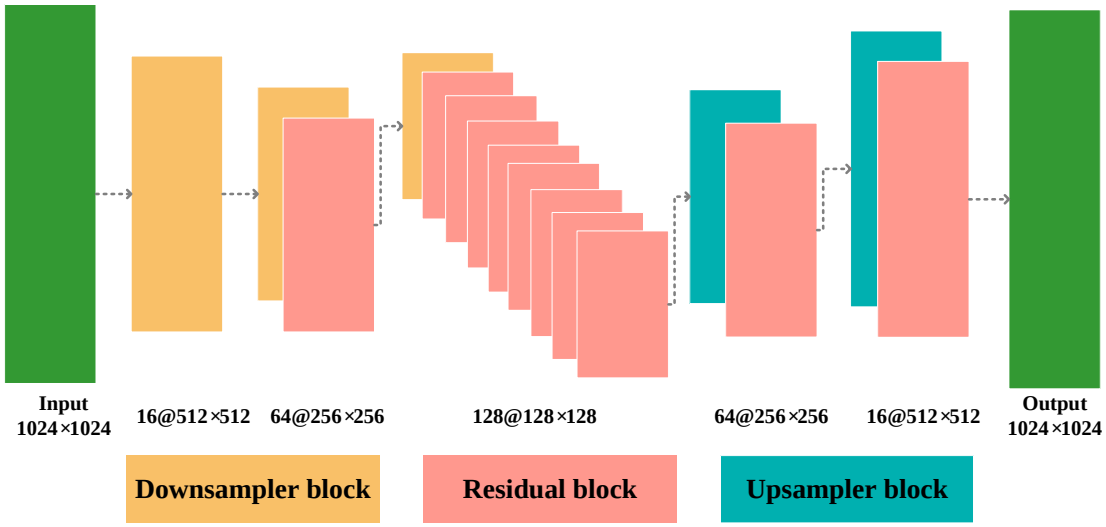

Figure 2: The overview of semantic segmentation network.*Number of feature maps at layers @ output resolution* is shown under each block. The network has three kinds of block. We follow an encoder-decoder architecture to avoid the need of using skip layers to refine the output. Furthermore, in consideration of simplifying the task, we only mark two kinds of information on the naturalistic image: the pupil and the iris.

As we know, Residual block consist of many stacked Residual Units and each unit can be expressed in a general form as $y_l = h(x_l) + F(x_l, W_l, x_{l+1} = f(y_l)$ where x_l and x_{l+1} are input and output of the $l - th$ unit, and F is a residual function. In $y_l = h(x_l) + F(x_l, W_l, x_{l+1} = f(y_l)$, $h(x_l) = x_l$ is an identity mapping and f is a ReLU function. We try to change the residual network structure makes the association between features stronger. By design the Residual block, we found the impact of our Residual block is twofold. First, the optimization is further eased (comparing with the baseline ResNet) because f is an identity mapping. Second, using BN as pre-activation improves regularization of the models. We follow an encoder-decoder architecture to avoid the need of using skip layers to refine the

output. Furthermore, in consideration of simplifying the task, we only mark two kinds of information on the naturalistic image: the pupil and the iris. However, many naturalistic images are influenced by light and other factors, and sometimes the pupil and the iris cannot be completely separated, to avoid "orphan semantic labels" that are only present in the input image, which the "orphan labels" usually are pupil region because of the outdoor illumination effect, we constrain the pupil semantic region to be set as the center of iris region. We have also observed that the segmentation does not need to be pixel accurate since eventually, the output is constrained by feature extraction network.

2.2. Generator G

We decompose the generator into two-subnetworks:G1 and G2. We term G1 as the global generator network and G2 as the local enhancer network. The generator is then given by the tuple $G = G1, G2$ as visualized in Fig. 4. The global generator network operates at a resolution of 297×297, and the local enhancer network outputs an image with a semantic layouts that is the output of the previous semantic segmentation network.

Our global generator is built on the architecture proposed by Johnson et al. [22], which has been proven successful for neural style transfer on images. It consists of 3 components: a convolutional front-end G1(F), a set of residual blocks G1(R) and a transposed convolutional back-end G1(B).

The local enhancer network also consists of 3 components: a convolutional front-end G2(F) , a set of residual blocks G2(R), and a transposed convolutional back-end G2(B). Different from the global generator network, a semantic label map is passed through the 3 components sequentially to output an image with instance segmentation informantion and the input to the residual block G2(R) is the element-wise sum of two feature maps: the output feature map of G2(F) , and the last feature map of the back-end of the global generator network G1(B). This helps integrating the global information from G1 to G2.

During training, we first train the global generator and then train the local enhancer in the order of their scale. We then jointly fine-tune all the networks together. We use this generator design to effectively aggregate global and local information for the image synthesis task.

2.3. Discriminator D

Realistic image synthesis poses a great challenge to the GAN discriminator design. To differentiate distribution naturalistic and synthesized images, the discriminator needs to have a large receptive field with instance segmentation information on global and local images. This would require either a deeper network or larger convolutional kernels. As both choices lead to an increased network capacity, overfitting would become more of a concern. Meanwhile, both choices require a larger memory footprint for training, which is already a scarce resource for realistic image generation. Inspired by Style Transfer, we proposed Discriminator D with novel loss function which is a pretrained VGG-19 (Simonyan and Zisserman (2014)) network and made some key modifications to the standard perception losses to keep the distribution of the naturalistic images and content of the synthetic images to the fullest extent. As Fig. 4 shows that instead of taking only RGB color channels into consideration, our network utilizes the representations of both color and semantic features

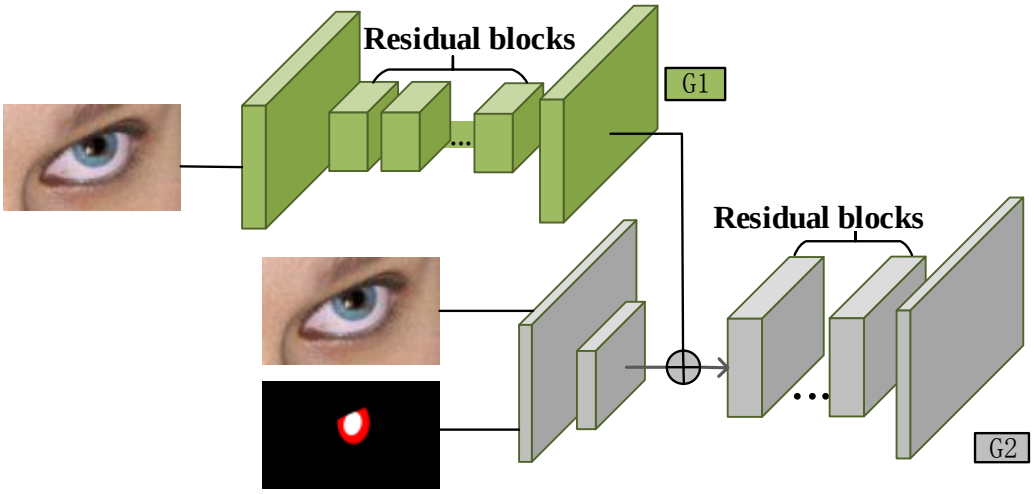

Figure 3: Network architecture of our generator. We first train a residual network G1 on global raw images. Then, another residual network G2 is appended to G1 and the two networks are trained jointly on local raw images with semantic layouts.

for style transfer. With the semantic features, we can address the spatial arrangement information and avoid the spatial configuration of the image being disrupted because of the style transformation.

2.3.1. Style reconstruction loss

Feature Gram matrices are effective at representing texture, because they capture global statistics across the image due to spatial averaging. Since textures are static, averaging over positions is required and makes Gram matrices fully blind to the global arrangement of objects inside the reference naturalistic image. So if we want to keep the global arrangement of objects, make the gram matrices more controllable to compute over the exact region of entire image, we need to add some texture information to the image. Luan et al. (2017) presented a method which added the masks to the input image as additional channels and augmented the neural style algorithm by concatenating the segmentation channels. Inspired by them, mask is added as the texture information to compute over the exact region of entire image. Thus the style loss can be denoted as:

$$\ell^l_{style} = \lambda_g \ell^l_{gs} + \lambda_l \ell^l_{ls} \tag{1}$$

$$\ell^l_{gs} = \sum_{c=1}^{C} \frac{1}{4N^2_{l,c}M^2_{l,c}} \sum_{ij} (G_l[O] - G_l[S])^2_{ij} \tag{2}$$

$$\ell^l_{ls} = \sum_{c=1}^{C} \frac{1}{4N^2_{l,c}M^2_{l,c}} \sum_{ij} (G_{l,c}[O] - G_{l,c}[S])^2_{ij} \tag{3}$$

where C is the number of channels in the semantic segmentation mask and l indicates the l-th convolutional layer of the deep convolutional neural network. Each layer with N_l distinct filters has N_l feature maps each of size M_l, where M_l is the height times the width of the feature map. So the responses in each layer l can be stored in a matrix $F[\cdot] \in R^{N_l \times M_l}$ where $F[\cdot]_{ij}$ is the activation of the i^{th} filter at position j in each layer l.

$$F_{l,c}[O] = F_l[O]S_{l,c}[I] \tag{4}$$

$$F_{l,c}[S] = F_l[S]S_{l,c}[S] \tag{5}$$

$$G_{l,c}[\cdot] = F_{l,c}[\cdot]F_{l,c}[\cdot]^T \tag{6}$$

$S_{l,c}[\cdot]$ is the segmentation mask in each layer l with the channel c. λ_g is the weight to configure layer preferences of global losses ℓ_{gs} which calculated between raw input image and features which was extracted by feature extraction network.λ_l is the weight to configure layer preferences of local losses ℓ_{ls} which calculated between input segmentation image and features which was extracted by feature extraction network with the input of segmentation image.

We formulate the style transfer objective as follows:

$$L_{total} = \sum_{l=1}^{L} \beta_l \ell_{style}^l \tag{7}$$

where L is the total number of convolutional layers and l indicates the l-th convolutional layer of the deep convolutional neural network. β_l is the weight to configure layer preferences. ℓ_{style} is the style loss(Eq.(4)). The advantage of this solution is that the requirement for mask is not too precise. It can not only retain the desired structural features, but also enhance the estimation of the pupil and iris information during the reconstruction of the naturalistic image style.

We now describe how we regularize this optimization scheme to preserve the structure of the input image and produce realistic but no distorted outputs. Our strategy is to express this constraint not on the output image directly but on the transformation that is applied to the input image. We name $Vc[O]$ the vectorized version $(N \times 1)$ of the output image O in channel c and define the following regularization term that penalizes outputs that are not well explained by a locally affine transform:

$$\ell_m = \sum_{c=1}^{3} Vc[O]^T Vc[O] \tag{8}$$

We formulate the realistic but no distorted style transfer objective by combining all components together:

$$L_{total} = \eta \sum_{l=1}^{L} \beta_l \ell_{style}^l + \vartheta \ell_m \tag{9}$$

where $\eta = 10^2$,$\vartheta = 10^4$

Our full objective combines both GAN loss ℓ_{GAN} and style tranfer loss D_{total} as:

$$min_G(\sum \ell_{GAN}(G, \ell^l) + \lambda \sum L_{total}) \tag{10}$$

where λ controls the importance of the two terms.

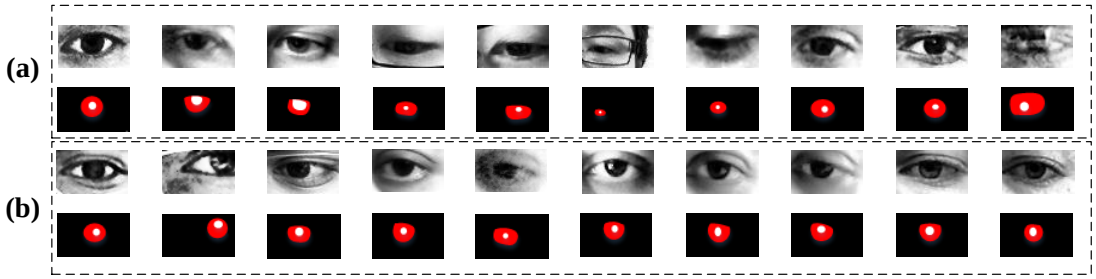

Figure 4: Coarse segmentation on MPIIGaze dataset.(a) represents images which come from training dataset and (b) represents images which come from testing dataset.Pupil region is labelled on white and iris region is red. We can observe that although testing dataset are influenced by light and other factors, the pupil and the iris can not be completely separated, proposed network can label the center of iris region to avoid "orphan semantic labels".

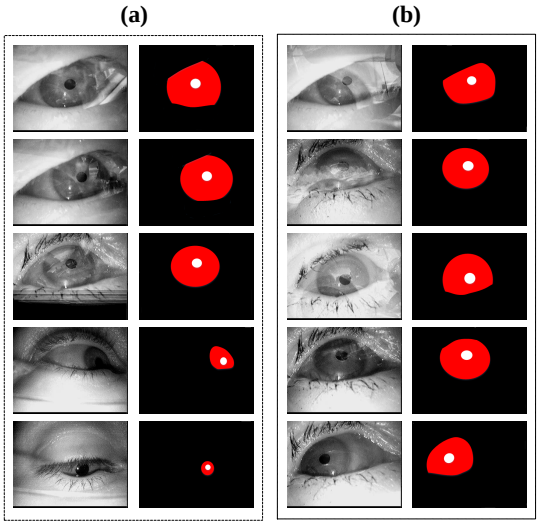

Figure 5: Coarse segmentation on LPW dataset.(a) represents images which come from training dataset and (b) represents images which come from testing dataset. For training coarse segmentation network, training dataset consists of 500 images from LPW datset and 1500 images from MPIIGaze datset which labelled on pupil and iris. Pupil region is labelled on white and iris region is red. we can observe that although the testing dataset under different illumination condition with training dataset, proposed network can achieve good results without "orphan semantic labels".

3. Experimental Results

3.1. Coarse Segmentation

Synthetic images can be segment easier than naturalistic images, thus we label the naturalistic image dataset for training a model which can segment pupil region and iris region effectively from naturalistic and synthetic image datasets.

Fig. 4 and Fig. 5 show the result of our coarse segmentation network on MPIIGaze dataset and LPW dataset respectively. (a) represents images which come from training dataset and (b) represents images which come from testing dataset. For training coarse segmentation network, training dataset consists of 500 images from LPW datset and 1500 images from MPIIGaze datset which labelled on pupil and iris. Pupil region is labelled on white and iris region is red. In the main paper we generate all comparison results using automatic coarse segmentation network.

From Fig. 5 we can observe that although the testing dataset under different illumination condition with training dataset, proposed network can achieve good results without "orphan semantic labels". Further more, we can observe that although testing dataset are influenced by light and other factors, the pupil and the iris can not be completely separated, proposed network can label the center of iris region to avoid "orphan semantic labels".

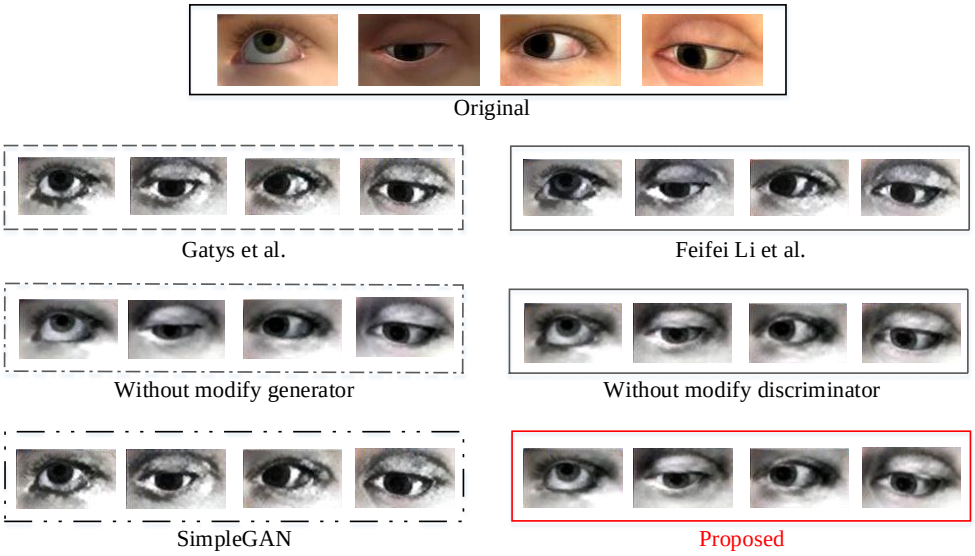

Figure 6: Example output of proposed method for UnityEyes gaze estimation dataset. The skin texture and the iris region in the refined synthetic images are qualitatively significantly more similar to the naturalistic images than to the synthetic images. What's more, comparing with generator and discriminator without modification , the distribution of pupil and iris regions are dramatically clear.

3.2. Qualitative Results

Besides that, we show the result of generator without modification and discriminator without modification. With all these five baseline methods, we show the result of two different datasets which are UnityEyes (Wood et al. (2016)) and SynthesEyes (Wood et al. (2015)).

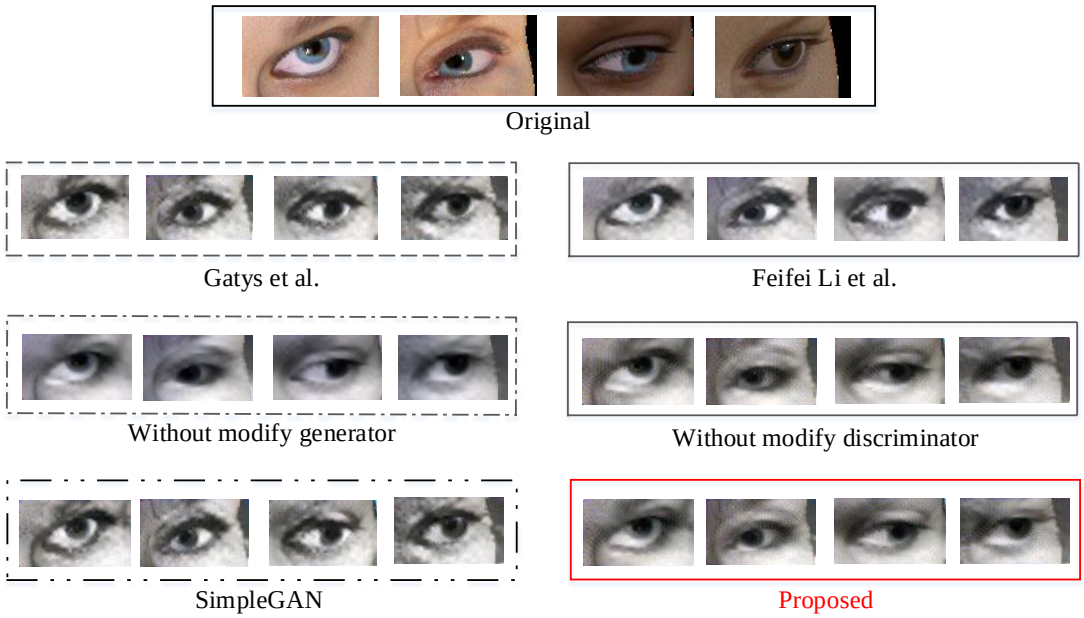

Figure 7: Example output of proposed method for SynthesEyes gaze estimation dataset. The skin texture and the iris region in the refined synthetic images are qualitatively significantly more similar to the naturalistic images than to the synthetic images. What's more, comparing with generator and discriminator without modification, the distribution of pupil and iris regions are dramatically clear.

As Fig.6 and Fig.7 we can see that if closely observed, it can be seen that none of these styles has similar gaze angle with naturalistic images. The skin texture and the iris region in the refined synthetic images are qualitatively significantly more similar to the naturalistic images than to the synthetic images. It can be observed that the proposed method is more similar with real conditions by light and achieves outstanding results above Gatys et al. (2015) and Feifei Li et al. Johnson et al. (2016). What's more, comparing with generator without modification and discriminator without modification, the distribution of pupil and iris regions are dramatically clear.

In order to validate the effectiveness of the proposed method, we compared it with available methods for several iterations in Fig.8 and Fig.9 on different datasets. "Iter" means the number of iteration. Because Shrivastava et al. (2016) is not stable so we only compare our method with Gatys et al. (2015) and Feifei Li et al. Johnson et al. (2016), we can see that after iteration for serval iterations, proposed method can achieve stable

distribution with less distortion, thus our result can be used to train a stable gaze estimator.

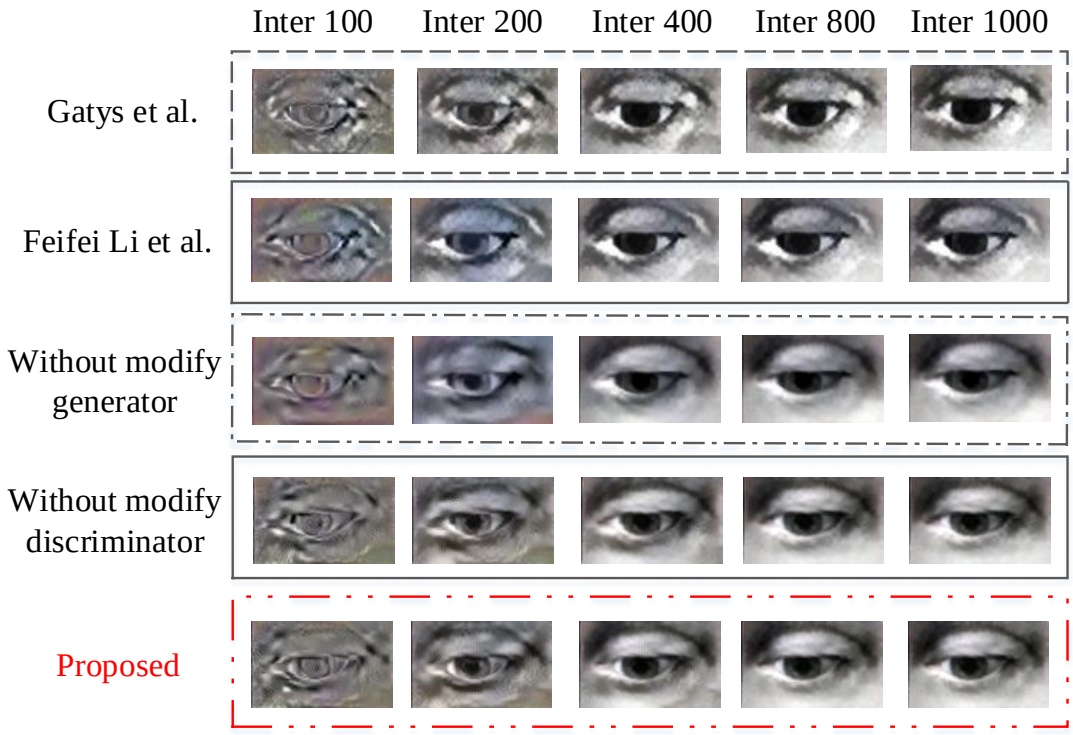

Figure 8: Example output of proposed method for UnityEyes gaze estimation dataset for several iterations. In order to validate the effectiveness of the proposed method, we compared it with available methods for several iterations. "Iter" means the number of iteration. Comparing our method with Gatys et al. (2015) and Feifei Li et al. Johnson et al. (2016), we can see that after iteration for serval iterations, proposed method can achieve stable distribution with less distortion.

3.3. Visual Turing Test

The most reliable known methodology for evaluating the realism of synthesized images is perceptual experiments with human observers. Such experiments yield quantitative results. There have also been attempted to design automatic measures that evaluate realism without humans in the loop. For example, Salimans et al. (2016) ran a pretrained image classification network on synthesized images and analyzed its predictions. We experimented with such automatic measures (for example using pretrained semantic segmentation networks) and found that they can all be fooled by augmenting any baseline to also optimize for the evaluated measure; the resulting images are not more realistic but score very highly (Salimans et al. (2016) Goodfellow et al. (2014)). Well-designed perceptual experiments with human observers are more reliable. We therefore use carefully designed perceptual experiments for quantitative evaluation.

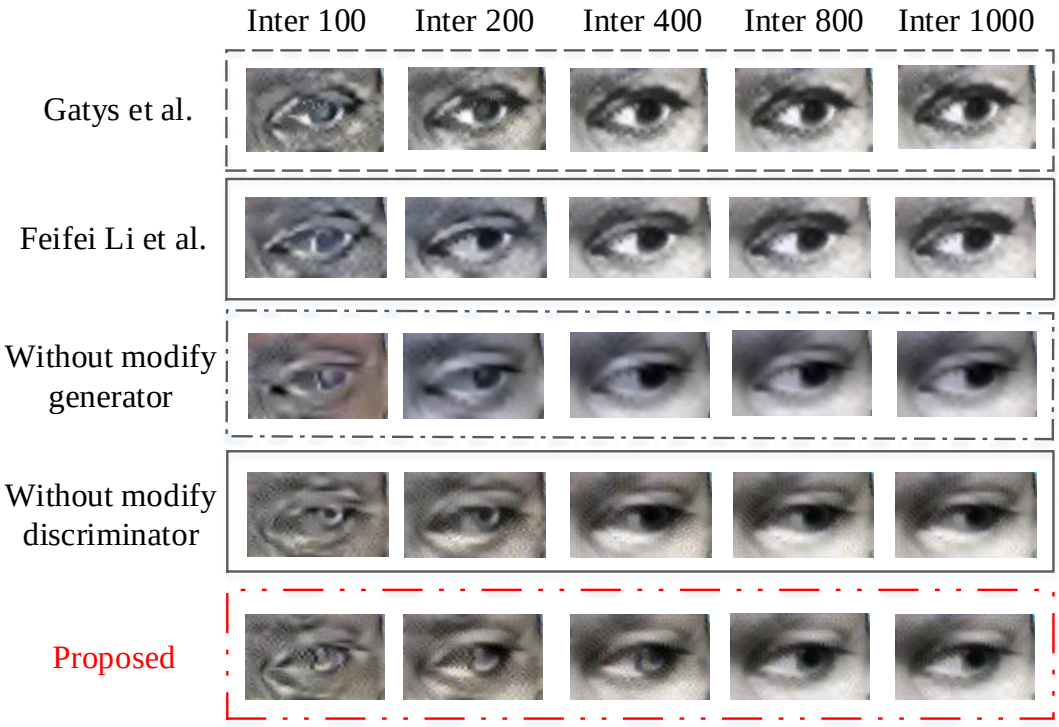

Figure 9: Example output of proposed method for SynthesEyes gaze estimation dataset for several iterations. In order to validate the effectiveness of the proposed method, we compared it with available methods for several iterations. "Iter" means the number of iteration. Comparing our method with Gatys et al. (2015) and Feifei Li et al. Johnson et al. (2016), we can see that after iteration for serval iterations, proposed method can achieve stable distribution with less distortion.

To quantitatively evaluate the visual quality of the refined images, we designed a user study where subjects were asked to classify images as naturalistic or refined synthetic. Each subject was shown a random selection of 200 naturalistic images and 200 refined images which were refined by proposed method in a random order, and was asked to label the images as either real or refined. The subjects found it very hard to tell the difference between the naturalistic images and the refined images. Table 1 shows the confusion matrix. In contrast, when testing on refined synthetic images which refined by Shrivastava et al. (2016) vs naturalistic images, we showed 50 naturalistic and 50 synthetic images per subject, and the subjects chose correctly 72 times out of 100 trials, which is significantly higher than ours.

Table 1: Results of the Visual Turing test user study for classifying real vs refined images by proposed method. Subjects were asked to distinguish between refined synthetic images (output from our method) and naturalistic images (from MPIIGaze). The average human classification accuracy was 50.6%, demonstrating that the automatically generated refined images are visually very hard to distinguish from naturalistic images.

	selected as real	selected as synthetic
real	1023	977
synthetic	1001	999

3.4. Appearance-based Gaze Estimation

To verify the effectiveness of the proposed method, we perform experiments to assess both the quality of our refined images and their suitability for appearance-based gaze estimation. We use COCO dataset to train the coarse model net. And few of images from MPIIGaze dataset are chosen as target images. The gaze estimation dataset consists of 28,332 synthetic images from eye gaze synthesizer UnityEyes-fine dataset, six subjects of UTview datset and 350,428 naturalistic images from the MPIIGaze dataset. For UTview Zhang et al. (2015), the data of subjects S0, S2, S3, S4, S6 and S8 in UTView are used as subject 1–6 in our dataset. In total, there are 144 (head pose) $\times$ 160 (gaze directions) $\times$ 6 (subjects) = 138,240 training samples and 8 (head pose) $\times$ 160 (gaze directions) $\times$ 6 (subjects) = 7680 testing samples.

We evaluate the ability of our method for appearance-based gaze estimation from naturalistic dataset and synthetic image dataset. ALR Lu et al. (2014), SVR Schneider et al. (2014), RF Sugano et al. (2014), convolution neural network Wood et al. (2015) and KNN Wood et al. (2015) are compared with our method as baseline methods. Similar to Wood et al. (2016), we train a convolution neural network (CNN) to predict the eye gaze direction. For RF training, pixel-wise data is employed to represent the original eye image by converting it to column vector, the number of trees during training is set to 20. For K-NN with UnityEyes refined images or UTview naturalistic images, considering that the computation cost increases with neighbor samples number, it can be found that a high-quality gaze estimator is obtained when the neighbor samples number is set to 50, which costs a shorter

operating time. A comparison to the state-of-the-art can be shown in Table.1. Training the CNN on the refined images outperforms the state-of-the-art on the part of MPIIGaze dataset. We observe that there is a large improvement in performance from training on the refined images and an significant improvement compared to the state-of-the-art.

Table 2: Comparison of our method to the state-of-the-art on the part of MPIIGaze dataset of real eyes which contains 350,428 images and UnityEyes dataset of synthetic images which contains 28,332 images of UnityEyes-fine dataset . The third column indicates whether the methods are trained on Real/Synthetic data. The error means eye gaze estimation error in degrees.

Method	Error	R/S
ALR Lu et al. (2014)	16.7	R
SVR Schneider et al. (2014)	16.6	R
RF Sugano et al. (2014)	15.4	R
CNN with UT Zhang et al. (2015)	13.2	R
K-NN with UT (ours)	8.9	R
CNN with UT (ours)	10.2	R
K-NN with Refined UnityEyes Wood et al. (2015)	10.2	S
CNN with Refined UnityEyes Wood et al. (2015)	11.5	S
CNN with Refined UnityEyes(SimGANs Shrivastava et al. (2016))	8.0	S
K-NN with Refined UnityEyes(ours)	8.3	S
CNN with Refined UnityEyes(ours)	7.7	S

4. Conclusion

We propose a coarse-to-fine eye synthesis method through adversarial training to speed up refining synthetic images with less unlabeled real data. We make several key modifications to the GANs to make the net become an efficient refine model net to improve the suitability of gaze estimation and make the image not distorted. We quantitatively evaluate the generated images by training models for gaze estimation. Comparing with the baseline methods, a large improvement in performance from training on the refined images is observed and the quantity of real data reduces by more than one order of magnitude.

References

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015. URL http://arxiv.org/abs/1508.06576.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The*

Netherlands, October 11-14, 2016, Proceedings, Part II, volume 9906 of *Lecture Notes in Computer Science*, pages 694–711. Springer, 2016. ISBN 978-3-319-46474-9. doi: 10.1007/978-3-319-46475-6_43. URL https://doi.org/10.1007/978-3-319-46475-6_43.

Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive linear regressionfor appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(10): 2033–2046, 2014. doi: 10.1109/TPAMI.2014.2313123. URL https://doi.org/10.1109/TPAMI.2014.2313123.

F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 00, pages 6997–7005, July 2017. doi: 10.1109/CVPR.2017.740. URL doi.ieeecomputersociety.org/10.1109/CVPR.2017.740.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016. URL http://arxiv.org/abs/1606.03498.

T. Schneider, B. Schauerte, and R. Stiefelhagen. Manifold alignment for person independent appearance-based gaze estimation. In *2014 22nd International Conference on Pattern Recognition*, pages 1167–1172, Aug 2014. doi: 10.1109/ICPR.2014.210.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. *CoRR*, abs/1612.07828, 2016. URL http://arxiv.org/abs/1612.07828.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL http://arxiv.org/abs/1409.1556.

Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, June 2014. doi: 10.1109/CVPR.2014.235.

Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3756–3764. IEEE Computer Society, 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.428. URL https://doi.org/10.1109/ICCV.2015.428.

Erroll Wood, Tadas Baltrusaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In Pernilla Qvarfordt and Dan Witzner Hansen, editors, *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, ETRA 2016, Charleston, SC, USA, March 14-17, 2016*, pages 131–138. ACM, 2016. ISBN 978-1-4503-4125-7. doi: 10.1145/2857491.2857492. URL http://doi.acm.org/10.1145/2857491.2857492.

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, June 2015. doi: 10.1109/CVPR.2015.7299081.