AOGAN: A generative adversarial network for screen space ambient occlusion

Lei Ren¹, Ying Song^{1,2,3} (✉)

© The Author(s) 2021.

Abstract Ambient occlusion (AO) is a widely-used real-time rendering technique which estimates light intensity on visible scene surfaces. Recently, a number of learning-based AO approaches have been proposed, which bring a new angle to solving screen space shading via a unified learning framework with competitive quality and speed. However, most such methods have high error for complex scenes or tend to ignore details. We propose an end-to-end generative adversarial network for the production of realistic AO, and explore the importance of perceptual loss in the generative model to AO accuracy. An attention mechanism is also described to improve the accuracy of details, whose effectiveness is demonstrated on a wide variety of scenes.

Keywords ambient occlusion (AO); attention mechanism; generative adversarial network (GAN); perceptual loss

1 Introduction

Ambient occlusion (AO) is a rendering technique widely used in real-time 3D applications. It simulates low-frequency shadows caused by ambient light due to object occlusions, which can effectively improve realism of appearance of rendered scenes. At present, there are mainly two groups of real-time AO approaches for dynamic scenes: screen-space ambient occlusion (SSAO) [1–4] and learning-based

methods [5–7], both of which are evaluated using screen-space information stored in a G-buffer. The former class are efficient, but often produce inaccurate results due to the empirical sampling strategy. The latter have gained much attention for their generality, but tend to miss details due to deficiencies in network structures or loss functions.

In a learning-based shading framework, a common idea is to train a neural network to learn from scene attributes to AO appearances as well as other screen-space shading effects. With powerful expressiveness and filter-sharing capacity, U-Net [8] and fully convolutional networks (FCNs) [9] have become the mainstream network structures for AO evaluation. Although these networks are representative, their predictive quality can still be improved [10, 11].

In this paper, AO evaluation is regarded as a generative learning task. For generative problems, deep learning-based methods can achieve high structural similarity (SSIM) [12] and peak signal-to-noise ratio (PSNR) [13]. Although pixel-wise errors in these works [7, 14, 15] are reduced, there are still noticeable differences in visual quality compared to the ground truth. For example, results often exhibit undesirable smoothness. To further optimize the model in feature space, perceptual metrics can be introduced in addition to traditional metrics on image tasks to achieve higher quality results [16, 17]. Since the emergence of generative adversarial networks (GANs) [18], a large number of researchers [19] have found that adversarial loss is helpful to obtain better perceptual quality. An additional way to improve the perceptual similarity between generated results and the ground truth is to use a pre-trained VGG-19 [20] network to evaluate perceptual loss. Although perceptual loss is popular in image transformation tasks, existing learning-based AO approaches only

1 Zhejiang Sci-Tech University, Hangzhou 310018, China. E-mail: L. Ren, DoerRenlei@outlook.com; Y. Song, ysong@zstu.edu.cn (✉).

2 2011 Collaborative Innovation Center for Garment Personal Customization of Zhejiang Province, China.

3 Key Lab of Silk and Culture Heritage and Products Design Digital Technology, Ministry of Culture and Tourism, China.

Manuscript received: 2021-05-13; accepted: 2021-07-13

use per-pixel metrics like PSNR and SSIM, where more attention is paid to pixel-wise differences rather than perceptual ones.

In AO results, human vision is sensitive to regions containing object edges. Self occlusion helps in understanding object shape, while mutual occlusion helps in understanding distances between objects. There are two advantages of introducing the attention mechanism. On the one hand, edge regions can be better integrated with global information through redistribution of weights, making evaluation of occlusion around complex geometries more accurate. On the other hand, for scenes with rich details, the learning process is more difficult and time-consuming; the attention mechanism can get the most significant feature subset through a dimensionality reduction process. For this purpose, we adopt the attention mechanism to learn AO results, which makes occlusion around geometric details more accurate.

In summary, our contributions are as follows:

- an end-to-end trainable generative adversarial network, AOGAN, which can generate high quality AO,
- use of perceptual loss to improve the overall visual quality of AO generation in complex scenes, and
- an attention mechanism to make evaluation of occlusion around complex geometries more accurate.

2 Related work

2.1 Screen-space ambient occlusion

SSAO was first proposed to overcome the huge computation and extra storage overhead of offline AO algorithms. Such methods are popular in modern 3D applications, but often produce artifacts such as white highlights at the edges of objects or dark halos around the contours of objects due to large depth differences in adjacent pixels in screen space. SSAO+ [4] is an extension of hemispheric sampling along surface normals, resolving the white highlight at the edge of objects but still producing dark halos. Horizon based ambient occlusion (HBAO) [22] approximates AO by finding the free horizon angle along each direction in a jittered sample pattern, and integrates over all directions. However, special care must be taken to avoid banding and noise. Alchemy

screen-space ambient obscurance (ASSAO) [23] uses a falloff function with user-specified parameters in the occlusion integral to approximate a more realistic result, but still does not directly deal with errors introduced by the screen space approximation. Vardis et al. [24] adopted an importance sampling strategy, and determined the key region for sampling by joint optimization over multiple views. Jimenez et al [25] proposed a new form of integration, GTAO, which can speed up the computation and improve accuracy. These methods shared a trait that the empirical sampling parameters or thresholds must be carefully chosen to avoid artifacts.

2.2 Deep learning for ambient occlusion

Recently, there has been a trend to use deep learning to solve rendering problems, such as neural network ambient occlusion (NNAO) [5] and DeepShading [7], which perform better than traditional algorithms with comparable rendering performance. In NNAO, the authors trained a multilayer perceptron (MLP) by using depth-normal pairs and the ground truth. Nalbach et al. [7] proved that CNNs are superior to MLPs because the former have larger receptive fields. However, because the loss function they adopted was pixel-level SSIM or L2 loss, they did not pay enough attention to edges or gaps between small objects, leading to local inaccuracies in occlusions (Fig. 1). Zhang et al. [21] improved rendering performance following the same idea as in Ref. [7]. They also contributed their own datasets. Although some scenes result in a high average SSIM, differences from the ground truth can still be seen (Fig. 1(c)).

2.3 Generative adversarial networks

GANs [18] have proved to be powerful in regenerating complex distributions. They can be classified into two categories according to the input data distribution, unconditional or conditional. Unconditional GANs generate real data from samples of general data distributions such as Gaussian noise; DC-GAN [26] is a good example which is designed primarily to produce general natural images. In conditional GANs [27], inputs related to the objective function are provided for the network to control the generated output. Thus, conditional GANs are very popular in conversion tasks, such as image translation [28], super-resolution [19], and neural rendering [29]. We model the AO task as a conversion problem, to be solved

Fig. 1 Comparison of ground truth (above) and CNN training results (below): (a) occlusion is brightened, (b) details are blurred, (c) global SSIM value is high but local differences are visible.

with a conditional GAN. With proper loss functions, it can seek differences both in feature space and at the pixel level, providing better visual quality.

2.4 Perceptual metrics

In the training process, the loss function plays an important role. Some theoretical and experimental [30, 31] analysis has proved that high-level features extracted from a pre-trained image classification network can retrieve perceptual information from input images. Specifically, in the pre-trained image classification network, the representation extracted from the hidden layer is helpful to interpret the semantics of input images. Researchers [19, 32] have found that the performance of image generation can be improved by using the Euclidean distance between high-level image features as a measure of depth perception similarity. In addition, works using perceptual metrics, like image style conversion [30, 33], novel view synthesis [31], etc., have also achieved convincing results.

2.5 Attention mechanism

In deep neural networks, the weights on connections can be regarded as resources, which are assumed to be evenly distributed at the beginning of training.

With an attention mechanism, they are reallocated according to the importance of objects: more important ones will get more resources. Recently, an attention module has been adopted in many tasks [34–37]. Vaswani et al. [37] first proposed to extract global dependency of the input by using such a module, for machine translation. At the same time, Zhang et al. [38] introduced attention mechanisms into GANs to get better generators. Inspired by these works, we also adopt an attention module in the generator to improve AO quality in the high-frequency region. Experiments in Section 4 show the effectiveness of this approach.

3 Method

3.1 Structure

The main structure of our proposed AOGAN is illustrated in Fig. 2. The inputs are normals and positions from the G-buffer. We send input into the generative network G to get an initial result, which then goes through the discriminant network D and the VGG network. The final difference is summarized into the loss function, and G and D are updated by back propagation.

Fig. 2 High-level overview of AOGAN. The inputs are normals and positions. Arrows represent the direction of operation at each layer of the network.

3.2 Joint loss function

3.2.1 Components

When training G , mean squared error (MSE) is commonly used as a loss function for AO problems. We introduce a new loss function, which takes semantic feature errors into account as well as pixel errors. The loss function $\mathcal{L}_G$ is represented as the weighted sum of the adversarial loss, the perceptual loss, and the content loss:

$$\mathcal{L}_G = \lambda_1 \ell_{\text{Gen}} + \lambda_2 \ell_{\text{per}}^j + \lambda_3 \ell_{\text{DSSIM}} \quad (1)$$

where λ_1 , λ_2 , and λ_3 are hyper-parameters which control the weights of each term.

3.2.2 Adversarial loss

The generative network G is the mapping from training data p_x to real-world data $G(x; \theta_g)$, where θ_g is the generator parameter. As shown in Eq. (2), the discriminant network D is trained to maximize the probability of getting the correct label for real-world data and $G(x; \theta_g)$, while the generative network G is trained to minimize $\log(1 - D(G(x)))$. These two networks are alternately trained by optimizing the value function $V(D, G)$.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_r} \log(D(x)) + \mathbb{E}_{x \sim p_g} \log(1 - D(G(x))) \quad (2)$$

In order to alleviate the collapsing mode problem, we modify the original GAN loss function following WGAN [39]. The generator loss function is

$$\ell_{\text{Gen}} = -\min_G \mathbb{E}_{x \sim p_g} [D(G(x))] \quad (3)$$

and the discriminator loss function is modified to be

$$\min_D \mathbb{E}_{x \sim p_g} [D(G(x))] - \mathbb{E}_{x \sim p_r} [D(x)] \quad (4)$$

where $x \sim p_r$ is the true data distribution and $x \sim p_g$ is the input data distribution.

3.2.3 Perceptual loss

As indicated above, pixel-wise losses and perceptual losses can be used simultaneously. Pixel-wise losses compensate for the differences in pixel space, but often produce blurry results [14, 40]. Perceptual losses extract differences between high-level features of images from well-trained classifiers. In order to reduce the distance between the generated value and the real output in feature space, we use the feature graph of the first several layers of a pre-trained VGG-19 [20] network, so as to obtain better perceptual quality. The perceptual loss is defined as

$$\ell_{\text{per}}^j = \|\Phi^j \text{VGG}(y) - \Phi^j \text{VGG}(G(x))\|_2 \quad (5)$$

where Φ represents the intermediate feature output of the VGG network, and j represents the layer. Experiments show that using 2.2 layer features achieves the best result.

3.2.4 Content loss

In order to further minimize the error between the estimated AO and the ground truth, we add SSIM as content loss function to G . A larger SSIM can accelerate convergence of training, but causes excessive smoothness. To avoid this, we set the weight to 0.5:

$$\ell_{\text{DSSIM}} = 0.5(1 - \ell_{\text{SSIM}}) \quad (6)$$

We use DSSIM as other loss functions should be minimized, while SSIM should be maximized.

3.3 Network structure

Figure 3 shows the generator structure using U-Net, which contains two parts, for down-sampling and up-sampling. A typical down-sampling block is composed of 4 layers: a convolutional layer, a BN layer, a LeakyRelu activation layer ($\alpha = 0.2$), and a pooling layer. An up-sampling block

Fig. 3 Skip connections in AOGAN. Left: down-sampling block components. Right: up-sampling block components. The attention module is added to the skip connection component.

is composed of 5 layers: a deconvolutional layer, an attention layer, a convolutional layer, a BN layer, and a LeakyRelu activation layer. The attention layer will be introduced in Section 3.4, which connects corresponding up and down sampling parts to further improve the generated AO visual effect. A deconvolutional layer is used in the up-sampling block instead of an interpolation layer because it has better generalization ability and can achieve better results at the expense of a little more computation. The discriminator network structure follows PatchGAN [28]. The image is convolved into N patches which can better represent local details, and the authenticity of the image is evaluated by averaging that of the patches.

3.4 Attention mechanism

Our attention module (AM) is designed to improve the accuracy of the generated result; see Fig. 4. The output of the attention module x' is evaluated as

$$x' = \omega^l (\sigma_2(W_c^T * \sigma_1(W_x^T * x^l + W_g^T * g))) * x^l \quad (7)$$

where l stands for the l -th layer, W_c^T , W_x^T , and W_g^T denote the convolution operations, σ_1 is the activation function, σ_2 is the sigmoid function, and ω^l is a weight. α^l and the feature map are multiplied

Fig. 4 Self-attention mechanism. x^l is the result of the PReLU layer in the down-sampling part, g is the deconvolution result for the corresponding layer, and α^l is the result of a series of operations of the same dimensions. W , H , and C represent the number of channels.

element-wise. The purpose is to highlight the image area and suppress irrelevant features. Adding this mechanism greatly improves the rendering of small and complex scenes, as will be shown in Fig. 6. For α^l of each layer, the weight ω^l is different and set to 1, 0.5, 0.25, 0.125, 0.0625 respectively: with increasing network level, the size of the feature graph decreases, and the influence of the weight of each layer increases, so the weights must be reduced.

4 Experiments

4.1 Datasets and training

The dataset was provided by Nalbach et al. [7] and Zhang et al. [21], and consists of 105,000 pairs of deferred shading G-buffer data for 27 scenes, with corresponding reference images. We used 90,000 pairs of images to train the network, 10,000 for verification, and the remaining 5000 for testing. Each pixel in the G-buffer contains a high-dimensional feature vector, in which positions and normals may be correlated and redundant, but these redundant attributes have almost no extra cost and can improve performance of the network to achieve AO effects. The size of the output AO image is 512×512 ; it can be generated from complex scenes in just a few milliseconds.

We implemented all of our models using the Tensorflow deep learning framework. Training was conducted on a GTX Titan-X GPU. Since the model is fully convolutional and is trained on image patches, it can be applied to images of arbitrary size. During training, G and D were trained alternately, and G updated D every 5 times on average. We set the decay to 0.9 for batch normalization and set $\lambda_1 = 0.001$, $\lambda_2 = 2$, $\lambda_3 = 1$ for $\mathcal{L}_G$. Learning uses the Adam optimizer, with a learning rate for all models of 10^{-4} , with a batch size of 8, and 30 epochs, these values empirically giving better results on validation. It took on average 160 hours to train an AOGAN network.

4.2 Comparison

To evaluate the performance of AOGAN, we conducted comparative experiments on our collected AO datasets. We calculated the SSIM and PSNR with generated AO and the ground truth to measure the error. However, we believe that pixel-wise quantitative evaluation is not necessarily the most effective metric, as we want to generate an AO that is more in line with the target, rather than being

exactly the same in content as the original image. We randomly selected multiple scenes as shown in Fig. 5. Even if numerical evaluation of SSIM of the results is almost the same as for the ground truth, there are obvious differences in visual effects. This difference comes from the loss function of the network, which is only the result after averaging the value of the image itself, instead of exploring high-level semantics.

Thus, we also take perceptual loss as one of the criteria for quantitative evaluation. Quantitative comparative results are shown in Table 1. Our approach achieves the best quality for several metrics. Experiments with IDs A2 and A3 in Table 1 show that although the SSIM value for DeepShading (DS) is higher, the perceptual loss value for HBAO is better

than that for DS. HBAO considers some basic features of the occlusion environment in viewing space, and these features lead to lower perceptual loss. This also explains why some visualization results of HBAO are better than those from DS.

We randomly selected four different scenes and present visual results of multiple algorithms for comparison in Fig. 6. Our results are superior to those of previous algorithms both in overall visual effect and in occlusion details along object edges and at corners. We attribute this to the nature of our approach which leads to more accurate generation and better image detail.

4.3 Ablation studies

4.3.1 Overview

Next, we validate the contribution of each component of the our approach through a thorough ablation study. To verify the effectiveness of our network architecture and loss function, we compare AOGAN with ablated versions:

- w/o $\ell_{\text{gen+per}}$: to fully consider the importance of perceived quality in the network, we remove

Table 1 Numerical evaluation of various baseline methods. Best scores are highlighted in bold

ID	Method	SSIM	PSNR	Perceptual	Run-time (ms)
A1	SSAO	0.695	17.68	0.310	14.4
A2	HBAO	0.751	19.38	0.186	8.5
A3	DS	0.803	21.85	0.231	12.5
A4	Ours	0.864	24.80	0.083	13.5

Fig. 5 Left to right: results from DeepShading, our method, and ground truth respectively. SSIM values are almost the same, but the results differ.

Fig. 6 Visual comparison of different methods. Left to right: (a) SSAO, (b) HBAO, (c) DeepShading, (d) ours, (e) ground Truth. Obvious differences are outlined in red.

the discriminator and ℓ_{per} from Eq. (1) ($\lambda_1 = 0$, $\lambda_2 = 0$).

- w/o ℓ_{per} : we remove ℓ_{per} from Eq. (1) ($\lambda_2 = 0$).
- w/o ℓ_{DSSIM} : we remove ℓ_{DSSIM} from Eq. (1) ($\lambda_3 = 0$).
- w/o AM: we remove the AM.

We train these ablated models under the same training conditions and datasets. Quantitative results of ablation studies are shown in Table 2, and example output of ablated methods are shown in Figs. 7 and 8.

4.3.2 Loss function

As we can see, our joint loss function achieves the best performance. The experiments for IDs B1 and B5 in Table 2 show that when the discriminator and VGG are removed, the SSIM value of B1 approximately reached the level when using joint loss, but the

Table 2 Results of ablation studies. The best scores are highlighted in bold

ID	Method	SSIM	PSNR	Perceptual
B1	w/o $\ell_{\text{gen+per}}$	0.851	24.76	0.228
B2	w/o ℓ_{per}	0.83	22.45	0.106
B3	w/o ℓ_{DSSIM}	0.81	23.62	0.098
B4	w/o AM	0.84	24.68	0.092
B5	Ours	0.864	24.80	0.083

perceptual loss value is the highest. We may doubt that the CNN can generate high-quality AO without using adversarial loss, while the joint loss is only slightly improved in overall detail. However, Fig. 7(a) shows that the addition of adversarial loss and perceptual quality significantly improves the effect in complex scenes, due to the fact that perceived loss makes it easier to generate high-level semantic

Fig. 7 Examples showing effects of ablation on loss function. Two locations are shown close-up to help see the differences.

Fig. 8 Examples showing effects of ablation on AM. Close-ups are shown at bottom left.

information. The reason for the lower SSIM value in B2 is that generator parameter updates are no longer dominated by SSIM value, but still show good visualization results: see Fig. 7(c).

In addition, in order to further study the importance of perceptual loss in the training process, we analyzed its effect by changing the value of λ_2 ; results are shown in Fig. 9. Making λ_2 smaller produces a smoother result; otherwise, it produces darker occlusion effects (see Fig. 9(c)): in the joint loss function, changing the value of λ_2 is equivalent to changing the proportion of λ_3 , which leads to a larger difference in pixel space. $\ell_{\text{per}} + \ell_{\text{DSSIM}}$ outperforms both ℓ_{per} and ℓ_{DSSIM} , which indicates that ℓ_{per} and ℓ_{DSSIM} promote each other.

4.3.3 Network modules

By comparing IDs B1, B4, and B5, we see that our complete network achieves the best performance. Since the generator itself cannot converge quickly, as shown in Fig. 7(a), the generated results in most complex scenes are relatively rough, which indicates that adversarial training can accelerate convergence of the generator. In order to further explore the role of the attention module, three network models

were trained under the same training conditions: DS, DS+AM, and DS_{deep}, where DS+AM adds a self-attention mechanism (AM) to the DeepShading network model and DS_{deep} deepens the network and increases the complexity through the residual network. We applied the rendering framework proposed by Zhang et al. [21] to further improve the network rendering time from 12.5 to 3.5 ms. Comparative results are provided in Table 3.

IDs C1–3 and C4&5 show that both deepening the network and adding AM can improve the accuracy of the model, but deepening the network negatively affects speed, which is unsuitable for real-time application. AM is based on calculating the covariance between each pixel and all other pixels to achieve a global estimate for each pixel; it can solve

Table 3 Evaluation of the attention mechanism for several networks. The best scores are highlighted in bold

ID	Method	SSIM	PSNR	Run-time (ms)
C1	DS	0.803	21.85	3.25
C2	DS+AM	0.831	24.76	3.4
C3	DS _{deep}	0.827	23.60	5.64
C4	GAN	0.84	24.68	3.20
C5	GAN+AM	0.864	24.80	3.5

Fig. 9 Effect of λ_2 on results. Obvious differences are outlined in red.

the problem of increasing the number of convolutional layers with a small computational cost. As shown in Fig. 8, this module can express the AO details of the object more clearly.

5 Conclusions

In this paper, we have proposed a new GAN-based ambient occlusion method, which can generate accurate screen-space ambient occlusion for complex scenes. At the same time, the improvement in quality did not come at the expense of running time, while the network structure and parameter specifications are similar to those in previous work [7]. Our running time is basically comparable and can achieve real-time performance. However, our work shares the same limitation as all screen-space solutions that the

G-buffers do not contain full information about the scene, which can cause artifacts on screen boundaries or culled scenes. One way to overcome this is to use an extended G-buffer or a multi-layer depth buffer as inputs to approximate more accurate AO. The proposed method is also suitable for other screen-space shading effects, which will be considered in our future work. Furthermore, we will also explore obtaining temporal consistency of AO with a GAN to avoid flicking between frames.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments. Ying Song was supported by the National Natural Science Foundation of China (No. 61602416), Shaoxing

Science and Technology Bureau Key Project (No. 2020B41006), and the Opening Fund (No. 2020WLB10) of the Key Laboratory of Silk Culture Heritage and Product Design Digital Technology.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

- [1] Mittring, M. Finding next gen: CryEngine 2. In: Proceedings of the ACM SIGGRAPH 2007 Courses, 97–121, 2007.
- [2] Shanmugam, P.; Arikan, O. Hardware accelerated ambient occlusion techniques on GPUs. In: Proceedings of the Symposium on Interactive 3D Graphics and Games, 73–80, 2007.
- [3] Bavoil, L.; Sainz, M. Screen space ambient occlusion. Available at <https://developer.download.nvidia.cn/SDK/10.5/direct3d/Source/ScreenSpaceAO/doc/ScreenSpaceAO.pdf>.
- [4] Fillion, D.; McNaughton, R. Effects & techniques. In: Proceedings of the ACM SIGGRAPH 2008 Games, 133–164, 2008.
- [5] Holden, D.; Saito, J.; Komura, T. Neural network ambient occlusion. In: Proceedings of the SIGGRAPH ASIA 2016 Technical Briefs, Article No. 9, 2016.
- [6] Erra, U.; Capece, N. F.; Agatiello, R. Ambient occlusion baking via a feed-forward neural network. In: Proceedings of the Eurographics - Short Papers, 13–16, 2017.
- [7] Nalbach, O.; Arabadzhyska, E.; Mehta, D.; Seidel, H. P.; Ritschel, T. Deep shading: Convolutional neural networks for screen space shading. *Computer Graphics Forum* Vol. 36, No. 4, 65–78, 2017.
- [8] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351*. Navab, N.; Hornegger, J.; Wells, W.; Frangi, A. Eds. Springer Cham, 234–241, 2015.
- [9] Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440, 2015.
- [10] Zhou, Z. W.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; Liang, J. M. UNet++: A nested U-net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Lecture Notes in Computer Science, Vol. 11045*. Stoyanov, D. et al. Eds. Springer Cham, 3–11, 2018.
- [11] Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B. et al. Attention U-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [12] Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* Vol. 13, No. 4, 600–612, 2004.
- [13] Avciabas, I.; Sankur, B.; Sayood, K. Statistical evaluation of image quality measures. *Journal of Electronic Imaging*, Vol. 11, No. 2, 206–223, 2002.
- [14] Dong, C.; Loy, C. C.; He, K. M.; Tang, X. O. Learning a deep convolutional network for image super-resolution. In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Vol. 8692*. Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 184–199, 2014.
- [15] Chen, G. Y.; Han, K.; Wong, K. Y. K. PS-FCN: A flexible learning framework for photometric stereo. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11213*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 3–19, 2018.
- [16] Blau, Y.; Michaeli, T. The perception-distortion tradeoff. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6228–6237, 2018.
- [17] Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 586–595, 2018.
- [18] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 2, 2672–2680, 2014.
- [19] Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. et al. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 105–114, 2017.
- [20] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

- [21] Zhang, D. J.; Xian, C. H.; Luo, G. L.; Xiong, Y. H.; Han, C. DeepAO: Efficient screen space ambient occlusion generation via deep network. *IEEE Access* Vol. 8, 64434–64441, 2020.
- [22] Bavoil, L.; Sainz, M.; Dimitrov, R. Image-space horizon-based ambient occlusion. In: *Proceedings of the ACM SIGGRAPH 2008 Talks*, Article No. 22, 2008.
- [23] McGuire, M.; Osman, B.; Bukowski, M.; Hennessy, P. The alchemy screen-space ambient obscurance algorithm. In: *Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics*, 25–32, 2011.
- [24] Vardis, K.; Papaioannou, G.; Gaitatzes, A. Multi-view ambient occlusion with importance sampling. In: *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 111–118, 2013.
- [25] Jimenez, J.; Wu, X.-C.; Pesce, A.; Jarabo, A. Practical real-time strategies for accurate indirect occlusion. In: *Proceedings of the SIGGRAPH 2016 Courses: Physically Based Shading in Theory and Practice*, 2016.
- [26] Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint* arXiv:1511.06434, 2015.
- [27] Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv preprint* arXiv:1411.1784, 2014.
- [28] Isola, P.; Zhu, J. Y.; Zhou, T. H.; Efros, A. A. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5967–5976, 2017.
- [29] Tewari, A.; Fried, O.; Thies, J.; Sitzmann, V.; Lombardi, S.; Sunkavalli, K.; Martin-Brualla, R.; Simon, T.; Saragih, J.; Nießner, M. et al. State of the art on neural rendering. *Computer Graphics Forum* Vol. 39, No. 2, 701–727, 2020.
- [30] Johnson, J.; Alahi, A.; Li, F. F. Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9906*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 694–711, 2016.
- [31] Park, E.; Yang, J. M.; Yumer, E.; Ceylan, D.; Berg, A. C. Transformation-grounded image generation network for novel 3D view synthesis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 702–711, 2017.
- [32] Dosovitskiy, A.; Brox, T. Generating images with perceptual similarity metrics based on deep networks. *arXiv preprint* arXiv:1602.02644, 2016.
- [33] Gatys, L.; Ecker, A.; Bethge, M. A neural algorithm of artistic style. *arXiv preprint* arXiv:1508.06576, 2015.
- [34] Lin, Z. H.; Feng, M. W.; Santos, C. N. D.; Yu, M.; Bengio, Y. A structured self-attentive sentence embedding. *arXiv preprint* arXiv:1703.03130, 2017.
- [35] Shen, T.; Zhou, T. Y.; Long, G. D.; Jiang, J.; Zhang, C. Q. DiSAN: Directional self-attention network for RNN/CNN-free language understanding. *arXiv preprint* arXiv:1709.04696, 2017.
- [36] Tang, J. H.; Hong, R. C.; Yan, S. C.; Chua, T. S.; Qi, G. J.; Jain, R. Image annotation by kNN-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology* Vol. 2, No. 2, Article No. 14, 2011.
- [37] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010, 2017.
- [38] Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In: *Proceedings of the 36th International Conference on Machine Learning*, 7354–7363, 2019.
- [39] Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv preprint* arXiv:1701.07875, 2017.
- [40] Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; Efros, A. A. Context encoders: Feature learning by inpainting. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2536–2544, 2016.

Lei Ren is a master student at the School of Information Science and Technology of Zhejiang Sci-Tech University. He got his B.S. degree from Hainan University in 2017. His research interests include deep learning and computer graphics.

Ying Song is an associate professor at the School of Information Science and Technology of Zhejiang Sci-Tech University. She obtained her B.S., M.S., and Ph.D. degrees in computer science and technology from Zhejiang University. She is a member of ACM SIGGRAPH academy. Her research interests are

appearance modeling, realistic rendering, and machine learning in computer graphics.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material

is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Other papers from this open access journal are available free of charge from <http://www.springer.com/journal/41095>. To submit a manuscript, please go to <https://www.editorialmanager.com/cvmj>.