MultiShadow: Shadow Synthesis for Multiple Virtual Objects

Ye Ding, Qi Wan, Ziyuan Liu, Jie Wang
School of Cyberspace Security, Dongguan University of Technology
dingye@dgut.edu.cn, wanqi719@126.com, 516807971@qq.com, 2111915024@dgut.edu.cn

***Abstract*—Image synthesis adds synthetic objects to existing images and makes them visually hard to distinguish. In order to make the synthetic objects more realistic, shadows are also required to be generated in the synthesis process. However, most existing shadow synthetic methods work poorly when there is no fixed lighting source or there are multiple objects that already existed in the original image. Hence in this paper, we propose a novel method called MultiShadow, which generates realistic shadows for virtual objects with better recognition of illumination information based on the existing objects in the original image. The experiments on benchmark datasets show that MultiShadow outperforms the state-of-the-art.**

***Index Terms*—image synthesis, shadow generation, adversarial attacks, fraud detection**

I. Introduction

Virtual objects are often added into an existing image in order to enhance the richness of content or restore the scene in virtual and augmented reality, as shown in Figure 1. In such application scenario, the synthesized image should contain few synthetic traces and emphasize the overall visual effects.

Previous image synthesis methods often focus on synthesizing the virtual object itself, such like adjusting the colour consistency, spatial layout and size of the virtual object. However, these methods only deal with the virtual object itself rather than blending in the background. In iconology, light and shadow can greatly enhance the three-dimensional sense of a picture and the corresponding objects [1]–[4]. Hence, shadow synthesis is an essential component of augmented reality, which aims to generate shadows for the virtual objects that have been added into the original image. Shadow synthesis is a novel research area and has been proven successful in increasing the realism of virtual objects [5].

Shadow synthesis is a challenging task majored because of the lack of information. 1) It is difficult to detect the direction of light source in the original image according to the perspective principles and geometric distortions of space. Without the light source, it is impossible to generate shadows in compliant with the other objects that already existed in the image. 2) Even if the light source is found, the three-dimensional information of the virtual object cannot be inferred through its two-dimensional projection. According to our observation, the lack of three-dimensional model of the virtual object is quite common in practise. 3) It is difficult to generate shadows for an added object in the image with multiple existing objects. Due to the variance in shape and position of existing objects, it is more difficult to decide the direction of light, and further influence the quality of generated shadows for added objects. In conclusion, a fined-grained shadow synthesis method should have the abilities of detecting the light source and generating shadows from two-dimensional projections [6], and be compatible with multiple existing objects.

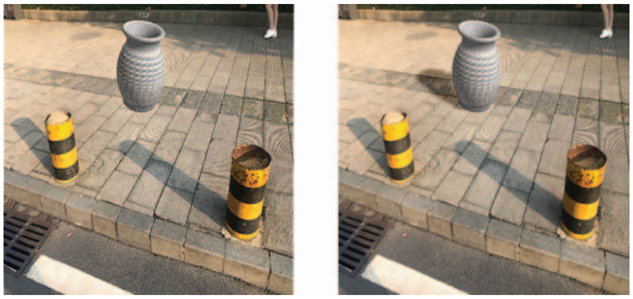


Fig. 1. An example of shadow synthesis using MultiShadow for the white pot as shown above in the image. The resulting shadow is realistic and in compliant with the direction of light source according to the other two objects.

In this paper, we propose a novel shadow synthesis method called MultiShadow as shown in Figure 1, where the generated shadows are well adapted into the original image. The contributions of this paper lie on the following aspects:

1) We propose a novel method called MultiShadow, which synthesizes shadows for the virtual objects that are added into an augmented image with multiple real objects that already exist.
2) We design an efficient two-stage model combining CNN and GAN together to detect and generate shadows in one network.
3) We have conducted thorough experiments on benchmark datasets, and the results show that MultiShadow outperforms the state-of-the-art.

II. Related Work

A. Shadow Detection

In order to generate shadows for virtual objects, it is necessary to recognize the existing shadows first. There are three major ways of shadow detection: 1) detect through environmental information with constant light. For instance, shadows could be segmented using spectral and geometric

978-1-6654-5968-6/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDIS55630.2022.00044

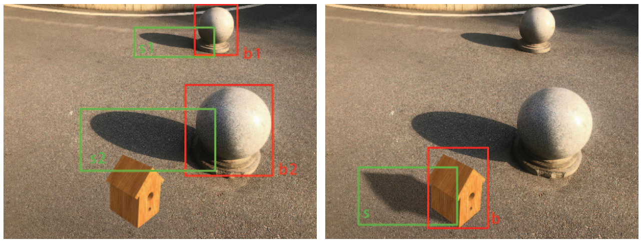
features in the scene [7], and the detected regions are verified as shadows based on colour variance and geometric properties. 2) Detect by extracting information from the edges of objects in the image, such like in outdoor consumer photographs [8], where the edges are filtered by a trained classifier. However, when the light source is far from the objects, shadows become blurry and unclear, thus the shaded areas are disturbed. In these scenarios, the above two methods are limited. Hence, a better way is 3) detecting by directly extracting features from shadows through deep neural networks. For example, using CNN (Convolutional Neural Networks) to capture the local structural information of shadows [9], [10], and using Patch-CNN [11] through a fully connected network to predict the probability of shadow edges [12]. In this paper, shadow detection methods are preliminaries that cannot be directly used to synthesize shadows.

B. Shadow Generation

Generating realistic shadows depends on whether the lighting information are provided. If illumination, reflectance, geometry and material properties [13] are already provided, it is trivial to generate shadows for virtual objects, and the rendering of geometric texture [14], object structure [15] and surface colour [16] is more important. However, in real life, there are often other objects with existing shadows in the background of an image, and the added virtual object must follow the existing lighting information. Hence, it is necessary to predict the illumination, reflectance, geometry, and material properties of the environment and the virtual object. It is possible to actively collect such information [17], [18] but time-consuming and labour-intensive. More advanced methods include: 1) Mask-ShadowGAN [19], which uses a mask to guide the generation of shadows and increase the number of object-shadow pairs in the dataset. However, it cannot handle complex backgrounds or specify target virtual objects. 2) ShadowGAN [20], which is based on both local and global adversarial discriminators to generate shadows. However, its lighting scenes are synthesized using a single point of light and it does not consider lighting conditions in real environments. 3) ARShadowGAN [5], which introduces an attention module to recognize real objects and corresponding shadows in the background of original image. Through learning the characteristics of real objects and corresponding shadows, it generates more realistic shadows for virtual objects. However, the performance of ARShadowGAN is limited on images with complex backgrounds.

III. MultiShadow

MultiShadow is a GAN-based model inspired by ARShadowGAN [5] and LISA [21], and it consists of a generator which generates shadows of virtual objects and a discriminator which verifies the quality of generated shadows. MultiShadow generates shadows without explicit lighting information and works for multiple objects simultaneously, which is different from previous works.



(a) Original image x (b) Ground truth image y

Fig. 2. An example of desired training image, where 1) there are multiple objects b_1, b_2 that already exist in the original image x; 2) shadows s_1 and s_2 are given for b_1 and b_2 respectively; 3) a virtual object b is inserted into the original image; and 4) the shadow s of the virtual object b is manually rendered, resulting ground truth image y.

A. Dataset

In this paper, the dataset for shadow generation model has the following requirements:

1) There multiple objects that already exist in the original image;
2) An object-shadow pair is given for each real object;
3) A virtual object is inserted to the original image; and
4) The shadow of the virtual object is manually rendered as ground truth.

An example is shown in Figure 2, where b is a 3D virtual object rendered by OpenGL, and the corresponding shadow of s should be synthesized based on the intensity and direction of light in the original image.

B. Framework

The design of MultiShadow is shown in Figure 3. MultiShadow is composed of a generator and a discriminator based on Generative Adversarial Network (GAN) [22]. The generator is responsible of generating shadows for the virtual objects, and the discriminator is responsible of verifying the quality of generated shadows. Through constant confrontation between the generator and the discriminator, synthesized shadows become more and more realistic during training.

The input of MultiShadow consists of the original image x and its corresponding object masks m. First, an attention encoder is introduced to fetch the objects and corresponding shadows that already exist in the background through the object decoder and shadow decoder, respectively. Then, the generator generates refined shadows $\hat{s}$ for the virtual objects in x. The above shadows are synthesized to the original image x resulting $\hat{y}$, and $\hat{y}$ is verified by the discriminator with confidence p. At last, p composes the loss of MultiShadow, and feeds back to the generator to improve the quality of $\hat{y}$ constantly.

Next, we will introduce the details of shadow generator, shadow discriminator, and loss function in Section III-C, III-D, and III-E, respectively.

C. Shadow Generator

Before generating the shadow $\overline{s}$ for the virtual object b in the original image x, an attention model is applied to

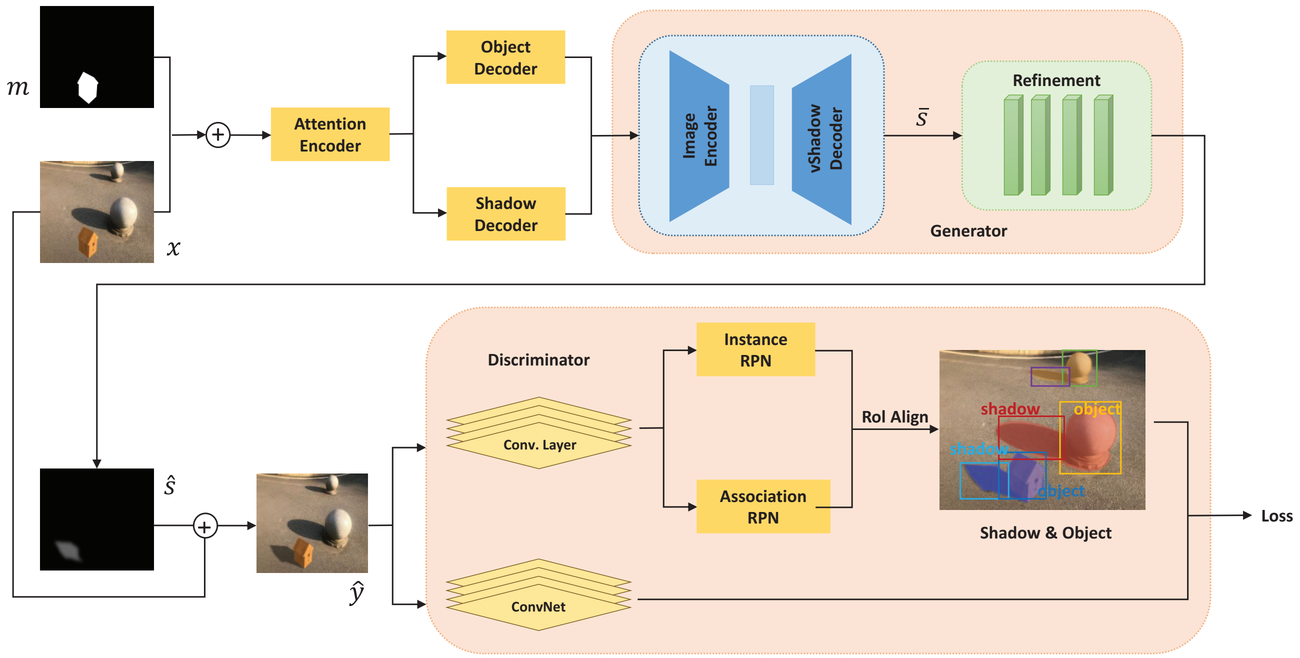


Fig. 3. The framework of MultiShadow.

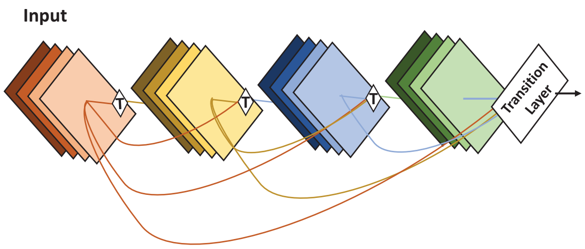


Fig. 4. The architecture of DenseNet. Each layer acquires features through non-linear transformations T such as BN, ReLU or Convolution. The features adopted in each layer is not only related to its adjacent layer, but also all the previous layers [23].

find those objects that already exist in the background. The attention model represents each region of the image as a matrix with elements ranging from 0 to 1, and features are extracted through a residual network. Then, these objects and their corresponding shadows are extracted from x. More specifically, two attention maps are obtained: 1) the attention map of the object occluders; and 2) the attention map of the corresponding shadows.

The shadow generator is composed of a U-Net [24] full convolutional network and a refinement network. The U-Net aims to generate rough shadows of virtual objects. More specifically in the U-Net, an image encoder is introduced to encode existing objects and their shadows, and a vShadow decoder is designed to decode a rough shadow $\bar{s}$ through the mask m of the virtual object b. The refinement network aims to optimize $\bar{s}$ and tries to find a refined shadow $\hat{s}$ for the object b. In this paper, we use DenseNet [23] as the refinement network, as shown in Figure 4. DenseNet introduces feature reuse mechanism to make the results more accurate.

D. Shadow Discriminator

The discriminator determines whether the generated shadows are realistic and feeds back the loss to the generator. In this paper, the discriminator of MultiShadow consists of two parts: 1) Continuous convolution with normalization and LeakReLU operations to generate feature maps for evaluation; and 2) a set of Region Proposal Networks (RPN) [25] to find areas with high probabilities to be objects and shadows in $\hat{y}$, and evaluated through RoI Align [26]. The loss of MultiShadow leverages the outputs of the above two parts by matching shadows and objects one by one. Failing the matching of shadow and object will result in a higher loss, and further enhance the performance of generator through feedbacks of average full-graph feature maps.

In this paper, we use two RPNs for the second part of the discriminator: 1) Instance RPN; and 2) Association RPN, as shown in Figure 5. RRN aims to identify the bounding boxes with targets. More specifically, Instance PRN identifies the object and shadow areas separately, whereas Association RPN identifies the whole area containing both the object and its corresponding shadow [21].

E. Loss Function

The loss of MultiShadow as shown in Figure 3 is denoted as L_G. In this paper, we combine 1) pixel-wise loss L_1; 2)

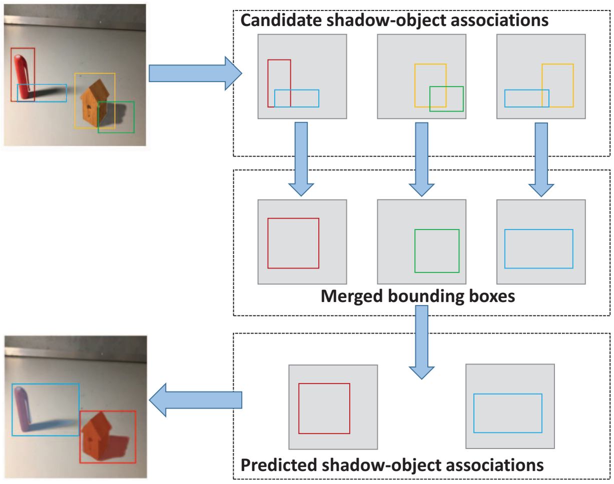

Candidate shadow-object associations
Merged bounding boxes
Predicted shadow-object associations

Fig. 5. RPN first recognises objects and shadows, then evaluate all the combinations of them to pick the associations with highest probabilities through RoI Align.

perceptual loss L_2; 3) cross entropy loss L_3; and 4) GAN loss L_4 together to calculate L_G. More specifically,

$$L_G = \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + \beta_4 L_4 \tag{1}$$

where β_1, β_2, β_3 and β_4 are hyper-parameters for L_1, L_1, L_3 and L_4, respectively.

L_1 is the pixel-wise loss that compares the synthetic image with its corresponding ground truth y. In this paper, L_1 compares both the rough image $\overline{y}$ and the refined image $\hat{y}$ with y, where $\overline{y}$ and $\hat{y}$ are synthetic by combining the original image x with the rough shadow $\overline{s}$ and the refined shadow $\hat{s}$. Formally,

$$\overline{y} = x + \overline{s} \tag{2}$$

$$\hat{y} = x + \hat{s} \tag{3}$$

and L_1 is defined as:

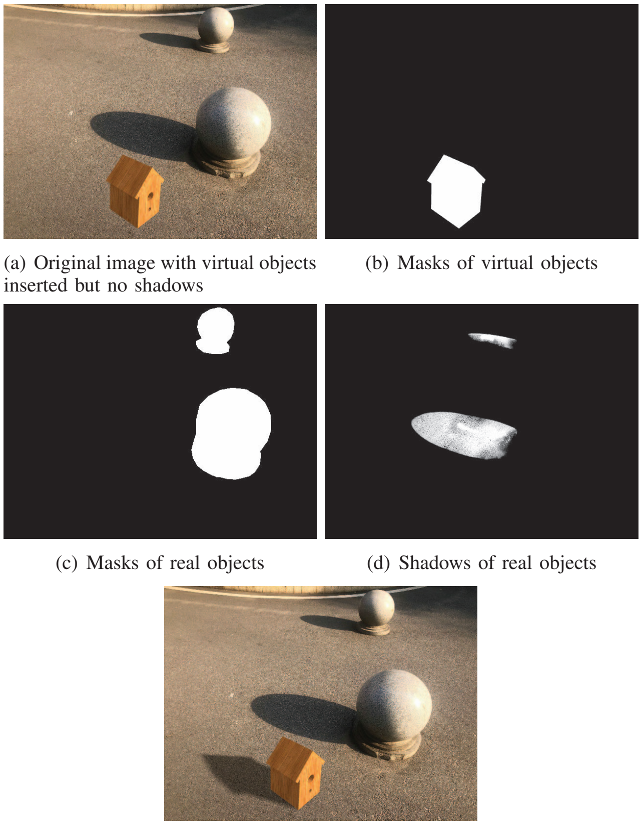
$$L_1 = \| y - \overline{y} \|_2^2 + \| y - \hat{y} \|_2^2 \tag{4}$$

L_2 is the perceptual loss which refers to the perceptual difference between the synthetic image and its corresponding ground truth. In this paper, we use a pre-trained VGG16 [27] model on ImageNet [28] to extract image features (denoted as V_y), and compare the mean squared errors with both the rough image $\overline{y}$ and the refined image $\hat{y}$. Formally,

$$L_2 = \mathrm{MSE}(V_y, V_{\overline{y}}) + \mathrm{MSE}(V_y, V_{\hat{y}}) \tag{5}$$

L_3 is cross entropy loss which evaluates the probability of successful predictions on object-shadow match as explained in Section III-D. $\hat{p}$ denotes the confidence of object-shadow match in the synthetic image, and p denotes the confidence of object-shadow match in the ground truth. Formally,

$$L_3 = -[p \log \hat{p} + (1-p)\log(1-\hat{p})] \tag{6}$$



(a) Original image with virtual objects inserted but no shadows

(b) Masks of virtual objects

(c) Masks of real objects

(d) Shadows of real objects

(e) Ground truth

Fig. 6. Envisioned data pairs of the ARShadowGAN dataset.

L_4 is the standard GAN loss that reflects the competition between generator and discriminator in MultiShadow. Formally,

$$L_4 = \log(D(x, m, y)) + \log(1 - D(x, m, \hat{y})) \tag{7}$$

where m is the mask of virtual object, and $D(\cdot)$ is the probability of identifying whether the image is real or fake (i.e., synthetic).

IV. Experiment

A. Dataset

Existing shadow datasets often have limits such as 1) the lack of corresponding object-shadow pairs [12], and 2) background only contains single object [29], [30]. According to the requirements of dataset as illustrated in Section III-A, object-shadow pairs and multiple background objects must be provided. Hence, in this paper, we use the dataset of ARShadowGAN [5] as the experimental dataset, which fulfils the above requirements. The ARShadowGAN dataset consists of background images taken by camera with multiple existing objects. The resolution of each image is 640×480 pixels. There are 13 different 3D virtual objects, where 4 of them are selected from a 3D scanning repository and 9 of them are selected from ShapeNet [31]. The envisioned data pairs

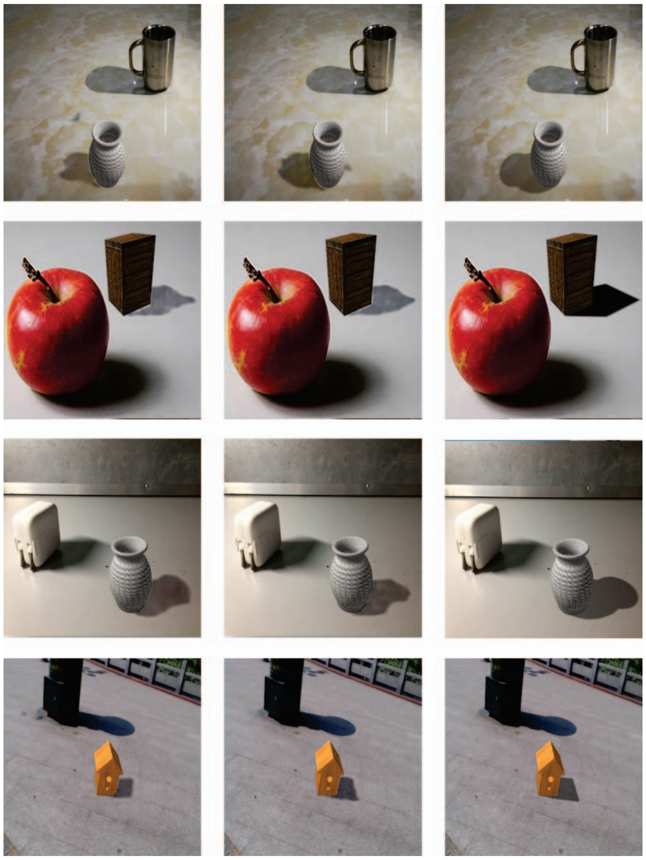


Fig. 7. Visual comparison of ARShadowGAN and MultiShadow on single-object backgrounds. Images from left to right are: ARShadowGAN, MultiShadow, and ground truth.

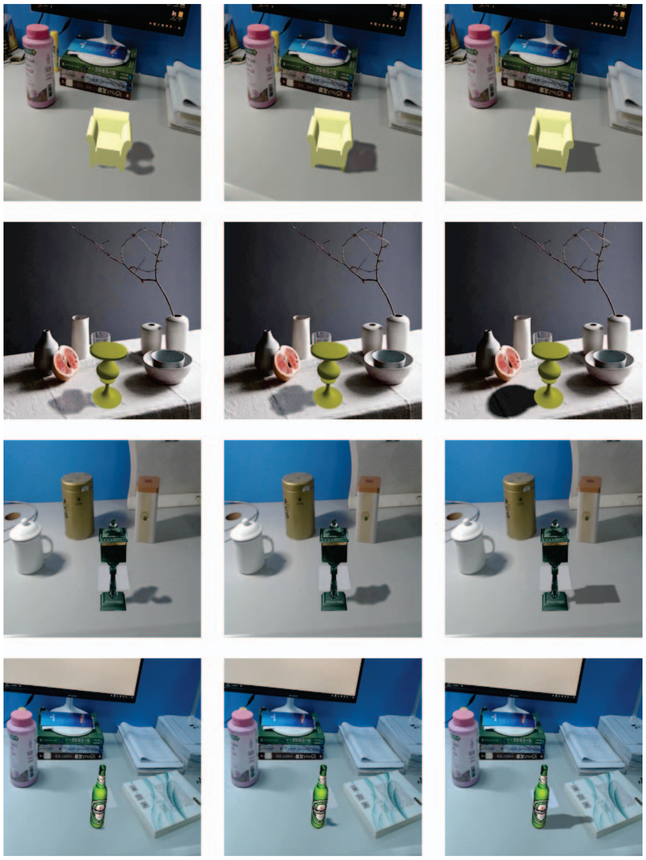


Fig. 8. Visual comparison of ARShadowGAN and MultiShadow on multi-object backgrounds. Images from left to right are: ARShadowGAN, MultiShadow and ground truth.

are shown in Figure 6, where each set of data pairs contains 5 images: 1) manually rendered original image with virtual objects inserted but no shadows; 2) masks of virtual objects; 3) masks of existing real objects; 4) shadows of existing real objects; and 5) manually rendered ground truth.

B. Evaluation Setup

We use PyTorch to implement MultiShadow. In the experiments, we set training epoch to 200, and use 5 up-sampling layers for the U-Net in the shadow generator. The hyper-parameters in the loss function of Equation (1) are set to: $\beta_1 = 10$, $\beta_2 = 1$, $\beta_3 = 0.011$, and $\beta_4 = 0.905$. The choice of these hyper-parameters are made through multiple experiments with best performance.

In this paper, we use the following three most commonly used evaluation metrics to evaluate the quality of synthesized image: 1) Root Mean Square Error (RMSE); 2) Structural Similarity Index (SSIM); and 3) Peak Signal-to-Noise Ratio (PSNR). For RMSE, the performance is better if the value is smaller. On the contrary, for SSIM and PSNR, the performance is better if the value is larger.

C. Experiment Result

As far as we know, ARShadowGAN is currently the state-of-the-art, and it outperforms all existing shadow generation algorithms [5]. Hence, in this paper, we only compare the performance of MultiShadow against ARShadowGAN with same training epochs. The overall performance and the performance only on multiple objects are shown in Table I. For the overall performance, MultiShadow is very close to ARShadowGAN on SSIM, and better on RMSE and PSNR, as shown in the left part of Table I. For the performance only on multiple objects, MultiShadow outperforms ARShadowGAN on all metrics. Hence, it proves that MultiShadow is clearly better for generating shadows on those images with multiple objects that already exist in the original image.

To clearly show the visual effects of MultiShadow, some examples of the synthesized images are shown in Figure 7 and 8 for single-object and multi-object backgrounds, respectively. According to Figure 7, MultiShadow performs similarly (and sometimes slightly better) to ARShadowGAN. However, accoriding to Figure 8, when there are multiple objects exist in the background, MultiShadow performs clearly better than ARShadowGAN.

TABLE I
COMPARISON OF ARSHADOWGAN AND MULTISHADOW ON OVERALL & MULTI-OBJECT PERFORMANCE

	Overall			Multi-Object Only		
	RMSE	SSIM	PSNR	RMSE	SSIM	PSNR
ARShadowGAN	7.53	**0.9767**	31.5	7.60	0.973	31.4
MultiShadow	**7.38**	0.9757	**31.7**	**6.99**	**0.974**	**32.2**

V. CONCLUSION

In this paper, we propose a novel shadow synthesis method called MultiShadow, which generates realistic shadows for virtual objects that are added to an existing image. The synthesized shadows are consistent with the lighting direction of the original image. MultiShadow works for images with complex backgrounds, especially when there are multiple objects that already exist in the original image. We have tested MultiShadow on benchmark datasets, and results show that MultiShadow outperforms the state-of-the-art.

ACKNOWLEDGEMENTS

This work is supported in part by the National Natural Science Foundation of China under grant no. 61976051, U19A2067, and U1811463.

REFERENCES

[1] J.-H. Kwon, S.-H. Nam, K. Yeom, and B.-J. You, “The use of shadows on real floor as a depth correction of stereoscopically visualized virtual objects,” in *IEEE International Symposium on Mixed and Augmented Reality - Media, Art, Social Science, Humanities and Design (ISMAR-MASH’D)*. IEEE, 2015, pp. 53–54.

[2] C. B. Madsen and F. Bertolini, “Outdoor illumination estimation for mobile augmented reality: Real-time analysis of shadow and lit surfaces to measure the daylight illumination,” in *International Conference on Computer Graphics Theory and Applications (GRAPP)*. Institute for Systems and Technologies of Information, Control and Communication, 2020, pp. 227–234.

[3] H. Wei, Y. Liu, G. Xing, Y. Zhang, and W. Huang, “Simulating shadow interactions for outdoor augmented reality with rgbd data,” *IEEE Access*, vol. 7, pp. 75 292–75 304, 2019.

[4] R. Brito, R. P. Biuk-Aghai, and S. Fong, “Gpu-based parallel shadow features generation at neural system for improving gait human activity recognition,” *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 12 293–12 308, 2021.

[5] D. Liu, C. Long, H. Zhang, H. Yu, X. Dong, and C. Xiao, “Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 1125–1134.

[7] E. Salvador, A. Cavallaro, and T. Ebrahimi, “Cast shadow segmentation using invariant color features,” *Computer Vision and Image Understanding*, vol. 95, no. 2, pp. 238–259, 2004.

[8] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan, “Detecting ground shadows in outdoor consumer photographs,” in *European Conference on Computer Vision (ECCV)*. Springer, 2010, pp. 322–335.

[9] L. Shen, T. Wee Chua, and K. Leman, “Shadow optimization from structured deep edge detection,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 2067–2074.

[10] W. Wu, K. Zhou, X.-D. Chen, and J.-H. Yong, “Light-weight shadow detection via gcn-based annotation strategy and knowledge distillation,” *Computer Vision and Image Understanding*, vol. 216, p. 103341, 2022.

[11] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, “Patch-based convolutional neural network for whole slide tissue image classification,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 2424–2433.

[12] T. F. Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, and D. Samaras, “Large-scale training of shadow detectors with noisily-annotated shadow examples,” in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 816–832.

[13] R. C. Yeoh and S. Z. Zhou, “Consistent real-time lighting for virtual objects in augmented reality,” in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2009, pp. 223–224.

[14] H. M. Kasem, K.-W. Hung, and J. Jiang, “Spatial transformer generative adversarial network for robust image super-resolution,” *IEEE Access*, vol. 7, pp. 182 993–183 009, 2019.

[15] S. van Steenkiste, K. Kurach, J. Schmidhuber, and S. Gelly, “Investigating object compositionality in generative adversarial networks,” *Neural Networks*, vol. 130, pp. 309–325, 2020.

[16] H. Wu, S. Zheng, J. Zhang, and K. Huang, “Gp-gan: Towards realistic high-resolution image blending,” in *International Conference on Multimedia (ACM)*. ACM, 2019, pp. 2487–2495.

[17] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sittig, and D. Forsyth, “Automatic scene inference for 3d object compositing,” *ACM Transactions on Graphics (TOG)*, vol. 33, no. 3, pp. 1–15, 2014.

[18] H. Weber, D. Prévost, and J.-F. Lalonde, “Learning to estimate indoor lighting from 3d objects,” in *International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 199–207.

[19] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng, “Mask-shadowgan: Learning to remove shadows from unpaired data,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 2472–2481.

[20] S. Zhang, R. Liang, and M. Wang, “Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks,” *Computational Visual Media*, vol. 5, no. 1, p. 8, 2019.

[21] T. Wang, X. Hu, Q. Wang, P.-A. Heng, and C.-W. Fu, “Instance shadow detection,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 1880–1889.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[23] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[24] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer International Publishing, 2015, pp. 234–241.

[25] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 91–99.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2961–2969.

[27] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[29] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau, “Deshadownet: A multi-context embedding deep network for shadow removal,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4067–4075.

[30] J. Wang, X. Li, and J. Yang, “Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 1788–1797.

[31] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, “Shapenet: An information-rich 3d model repository,” *arXiv:1512.03012*, 2015.