An Assessment of Shadow Generation by GAN with Depth Images on Non-Planar Backgrounds

Kaito Toyama and Maki Sugimoto
Graduate School of Open Environmental Science, Keio University,
3-14-1 Hiyoshi Kohoku-ku Yokohama, Kanagawa, Japan
{*kaito.t-0504, maki.sugimoto*}*@keio.jp*

Keywords: Mixed Reality, Shadow Generation, Virtual Object.

Abstract: We propose the use of a Generative Adversarial Network (GAN) with depth images to generate shadows for virtual objects in mixed reality environments. This approach improves the accuracy of shadow generation process by aligning shadows with non-planar geometries. While traditional methods require detailed lighting and geometry data, recent research has emerged that generates shadows by learning from the image itself, even when such conditions are not fully known. However, these studies are limited to projecting shadows only onto the ground: a planar geometry. Our dataset used for training the GAN, includes depth images allows natural shadow generation in complex environments.

1 INTRODUCTION

Shadow generation for virtual objects in Mixed Reality involves synthesizing shadows that align with real-world light sources and geometry when combining virtual objects with real-world images (Schrater and Kersten, 2000). Shadows are crucial for depth perception , which enhances the realism of virtual experiences (Hoffman et al., 1998; Chrysanthakopoulou and Moustakas, 2024).

Typically, shadow generation algorithms require information about real-world lighting, geometry, and the relative position and shape of virtual objects. However, obtaining this information is often challenging due to factors like occluded light sources and complex geometries.

To address this, machine learning approaches have been proposed. Liu et al. developed ARShadowGAN (Liu et al., 2020), a method that uses a Generative Adversarial Network (GAN) to generate virtual shadows that resemble real-world shadows by learning from the real-world objects and their shadows. This method successfully generates realistic shadows using only two inputs: a composite image without shadows and a mask of the virtual object. Additionally, Qi et al. introduced HAU-Net and IF-Net to account for lighting conditions and shadow shape, improving the naturalness of generated shadows (Meng et al., 2023).

However, these methods are limited by their assumption that shadows are projected onto a flat surface, making them less effective for complex geometries. In this study, we utilize depth images to generate natural shadows onto non-planar geometries. Depth images provide information about the distance between a camera and the target geometory, which can be considered as a key of shadow generation onto non-planar geometries. By incorporating depth images as an extension of the existing GAN framework, this paper investigates whether shadows considering depth information can be generated onto non-planar geometries.

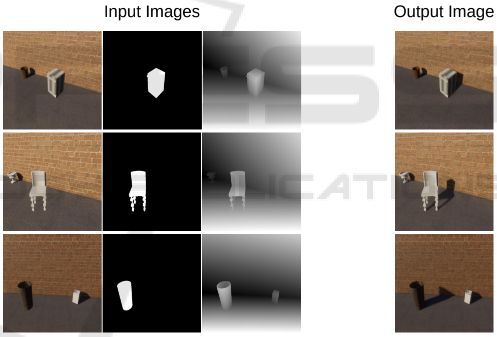
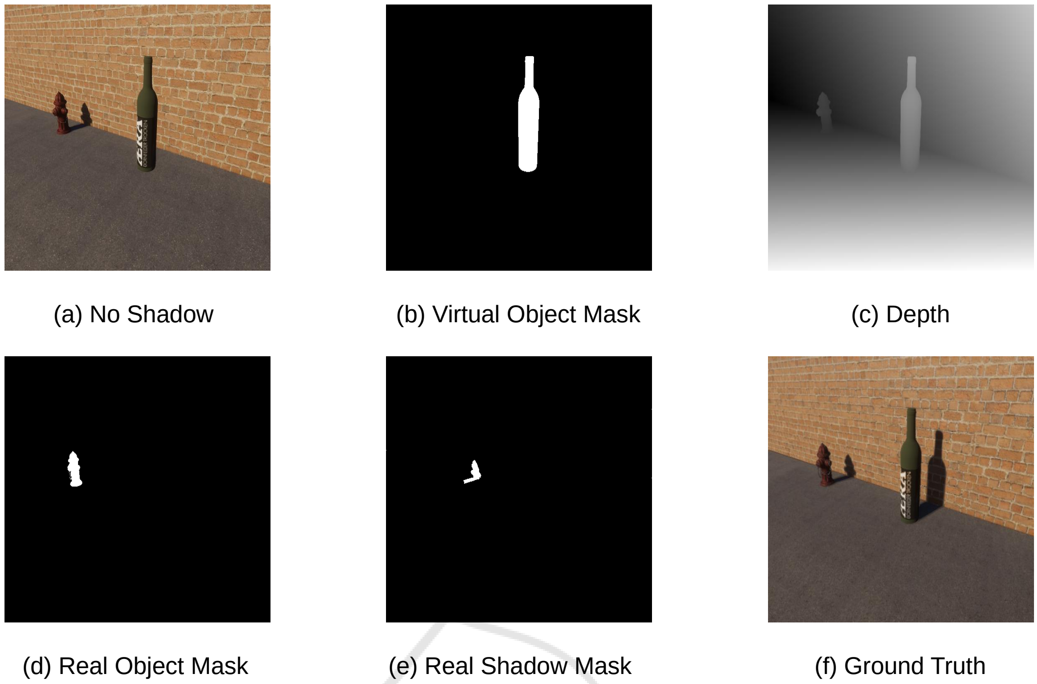


Figure 1: Left figures indicate the input images. Each image includes a shadowless virtual object with depth images. Right images indicate generated shadows by the virtual objects on non-planar backgrounds.

Toyama, K. and Sugimoto, M.
An Assessment of Shadow Generation by GAN with Depth Images on Non-Planar Backgrounds.
DOI: 10.5220/0013188200003912
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 135-142
ISBN: 978-989-758-728-3; ISSN: 2184-4321
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.



(a) No Shadow (b) Virtual Object Mask (c) Depth

(d) Real Object Mask (e) Real Shadow Mask (f) Ground Truth

Figure 2: The structure of the dataset. It consists of (a): No Shadow image, (b): Virtual Object Mask, (c): Depth image, (d): Real Object Mask, (e): Real Shadow Mask and (f): Ground Truth image.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

2 RELATED WORK

Shadow generation for virtual objects in MR environments can be achieved using algorithms when lighting conditions and real-world geometry are known. However, when these conditions are not fully available, shadow generation becomes challenging. To address this issue, machine learning-based methods have been developed.

Liu et al. proposed a method that integrates an Attention Block and a Shadow Generator Block into a GAN. This approach takes a real-world image with a synthesized virtual object and its corresponding mask as input. The Attention Block estimates the positions of real-world objects and their shadows, while the Shadow Generator Block uses this information along with the input images to generate shadows for the virtual object. The generated shadow is then combined with the real-world image to create a composite image, which is fed into the Discriminator. Both the Attention Block and Shadow Generator Block utilize U-Net (Ronneberger et al., 2015) for enhanced segmentation accuracy.

Meng et al. addressed the issue of unrealistic shadows in complex lighting environments by employing HAU-Net and IFNet. HAU-Net captures the interactions between foreground objects and the background in both spatial and channel dimensions, predicting realistic shadow shapes while considering background lighting conditions. IFNet uses Exposure Fusion to generate and merge multiple exposure images, which helps enhance the realism of shadows under varying lighting conditions. This process adapts to changes in background lighting, adjusting shadow intensity and shape, resulting in more natural and realistic shadows in the final composite image.

Sheng et al. introduced PixHt-Lab (Sheng et al., 2023), a system that leverages a pixel height-based representation to generate realistic lighting effects for image compositing, such as soft shadows and reflections. Unlike methods that assume shadows are projected onto a ground plane, PixHt-Lab maps the 2.5D pixel height representation to a 3D space, enabling the reconstruction of both foreground and background geometries. This approach significantly enhances soft shadow quality on general shadow receivers like walls and curved surfaces by incorporating 3D-aware buffer channels. Their neural renderer, SSG++, utilizes these buffer channels to guide soft shadow generation, addressing limitations in shadow realism and providing more control over lighting effects. While PixHt-Lab focuses on 2D image compositing, its use of geometry-aware data structures to guide lighting effects aligns closely with our approach to generating shadows for virtual objects on non-planar backgrounds.

To further improve the realism and quality of generated images, incorporating depth images into the generation process has proven effective. Depth images provide essential information about the distance from each pixel to the object surface, offering valuable insights into the shape of objects in RGB images. Qi et al. proposed a method that integrates depth images into the image generation process, significantly enhancing the naturalness and overall quality of the generated images. Their approach combines semantic labels with depth maps using a multi-conditional semantic image generation method. The quality improvement is achieved through the Multi-scale Feature Extraction and Information Fusion Module (MEIF) and the Multi-scale Channel Attention Module (MCA). MEIF leverages both depth information and semantic labels, using a pyramid-shaped feature extraction mechanism to capture both global and local details from the depth image, thus enhancing the feature maps derived from depth information. MCA aligns features across different scales by learning the correlation between feature map channels at varying scales, ensuring consistency and coherence in the generated images.

3 DATASET

We constructed a dataset which consists of No shadow images, Real object masks, Real shadow masks, Virtual object masks, Depth images and Truth shadow images, to train the network model. The image dataset used in this study Figure 2 was rendered using Unity 2021.3.16f1, with objects within the images sourced from ShapeNet (Chang et al., 2015) and BlenderCity (Couturier, 2023). Depth images were obtained by calculating the depth from the camera coordinates using the Z-buffer method. To generate mask images for background and virtual objects, we made non-background objects invisible and applied the Z-buffer method, where only the object portions had pixel values greater than 0. These images were then binarized to create mask images. The shadow mask images were generated by turning off the Cast Shadow setting for the objects and then taking the difference from the original image, followed by binarization. We generated 3,600 images by combining 20 background types, 3 viewpoints, and 60 virtual objects. The dataset will be published at https://github.com/YaMaKaTsu5004/DepthShadowGAN_dataset].

4 METHODOLOGY

Figure 3 shows the architecture of the network model used in this study. We extended the network model of ARShadowGAN (Liu et al., 2020) to generate shadows with depth information. Broadly, it consists of three main components: the Attention Block, the Shadow Generator Block, and the Discriminator Block.

4.1 Attention Block

In the Attention Block, segmentation of real-world objects and their shadows, which serve as input for the Shadow Generator Block, is performed. The input images consist of a composite image without shadows, a mask image of the virtual object, and a real-world depth image. The encoder uses ResNet (Residual Neural Network) (He et al., 2016), and U-Net (Ronneberger et al., 2015) is employed to learn the segmentation of real-world objects and their shadows.

4.2 Shadow Generator Block

The Shadow Generator Block is responsible for generating shadows for virtual objects. The input images consist of a composite image without shadows, a mask image of the virtual object, a real-world depth image, and the mask images of real-world objects and their shadows generated by the Attention Block. ResNet18 (He et al., 2016) is used as the encoder, and U-Net (Ronneberger et al., 2015) is employed to generate a rough shadow. The generated shadow is then refined through a Refinement process to make it more natural.

4.3 Discriminator

The Discriminator distinguishes whether the generated virtual shadow is plausible, thereby advancing the training of the generator. In this study, the Discriminator model adopts the PatchGAN method (Isola et al., 2017). PatchGAN consists of three consecutive convolutional layers, followed by Batch Normalize 2D and Leaky ReLU. Next, a convolution generates the final feature map, which is activated using a sigmoid function. The final output of the Discriminator is the global average pooling of the activated final feature map. In this study, the Discriminator takes as input the concatenation of the virtual object shadow, the virtual object mask, and the image containing the shadow of the virtual object.

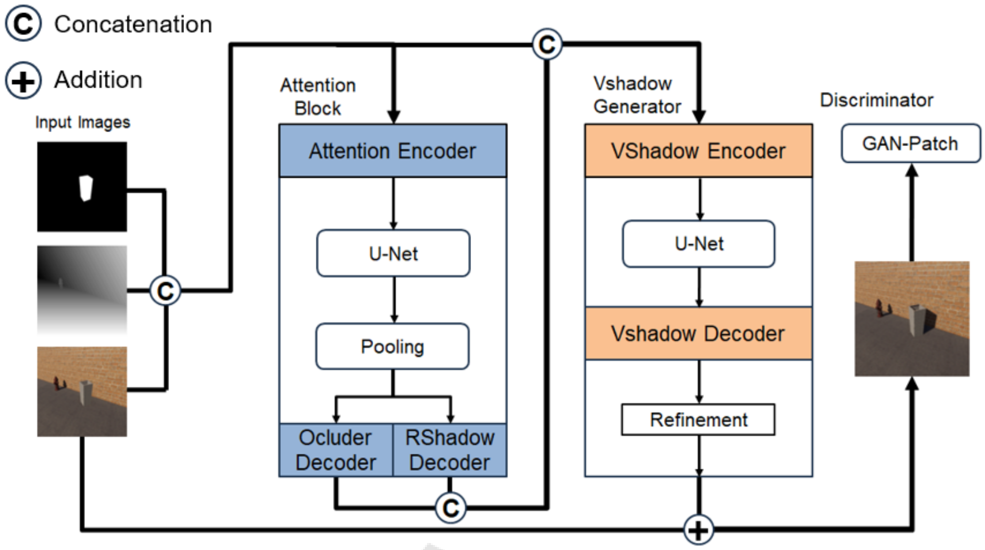


Figure 3: The architecture used in this study. The Attention Block focuses on real objects and their shadows, while the Virtual Shadow Generator generates shadows for virtual objects. The input images consist of a Virtual Object Mask, a Depth image, and No Shadow images. During training, the generated images are evaluated for authenticity by the Discriminator.

4.4 Loss Function

The loss function L_{attn} used in the AttentionBlock is defined using a squared loss as follows:

$$L_{\text{attn}} = \left\|A_{\text{obj}}(x,m,d) - M_{r_{\text{obj}}}\right\|_2^2 + \left\|A_{\text{shadow}}(x,m,d) - M_{r_{\text{shadow}}}\right\|_2^2 \quad (1)$$

where A_{obj} is the predicted mask image of the background object, and $M_{r_{\text{obj}}}$ is the ground truth mask image of the background object. Similarly, A_{shadow} is the predicted mask image of the background object's shadow, and $M_{r_{\text{shadow}}}$ is the ground truth mask image of the background object's shadow. The inputs (x,m,d) represent the background image, the mask image of the virtual object, and the depth image, respectively.

The loss function L_{gen} for the ShadowGeneratorBlock is defined as the weighted sum of three different terms:

$$L_{\text{gen}} = \beta_1 L_2 + \beta_2 L_{\text{per}} + \beta_3 L_{\text{adv}} \quad (2)$$

Here, L_2 represents the squared loss, which calculates the squared error between the predicted and true values. This is used to measure the error by treating the real composite image and the generated composite image as a regression problem. In the model of this study, refinement is performed to adjust the coarse shadows into more natural shadows. Therefore, the loss function calculates the squared loss both before and after the refinement, and their sum is defined as L_2. The output before refinement is:

$$\bar{y} = x + G(x,m,A_{\text{obj}},A_{\text{shadow}}) \quad (3)$$

and the output after refinement is:

$$\hat{y} = x + R(G(x,m,A_{\text{obj}},A_{\text{shadow}})) \quad (4)$$

The final L_2 loss is defined as follows:

$$L_2 = \|y-\bar{y}\|_2^2 + \|y-\hat{y}\|_2^2 \quad (5)$$

L_{per} is the perceptual loss (Johnson et al., 2016), which emphasizes high-level feature and structural similarity of the images. In this study, we use VGG16 (Simonyan and Zisserman, 2015) pre-trained on ImageNet (Deng et al., 2009) to extract features. This function is defined as follows:

$$L_{\text{per}} = \text{MSE}(V_y, V_{\bar{y}}) + \text{MSE}(V_y, V_{\hat{y}}) \quad (6)$$

where MSE is the mean squared error, and V_y represents the feature map extracted by the pre-trained VGG16. This calculation compares the feature maps at the intermediate layers to compute the loss.

The loss function for the Discriminator is defined as follows:

$$L_{\text{adv}} = \log(D(x,m,y)) + \log(1 - D(x,m,\hat{y})) \quad (7)$$

Here, D represents the probability that the image is real. During GAN training, the Discriminator tries to maximize $L_a dv$, while the Generator tries to minimize it.

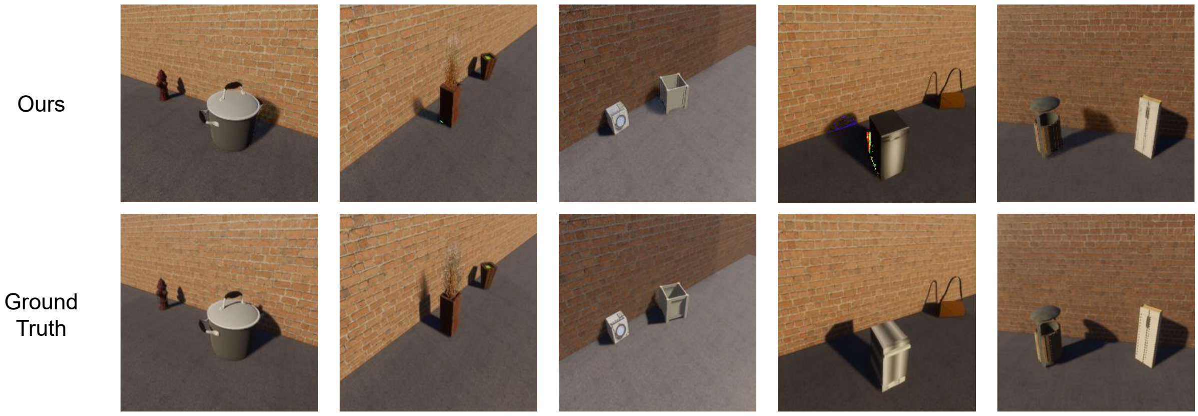


Figure 4: A comparison of our method and the ground truth.

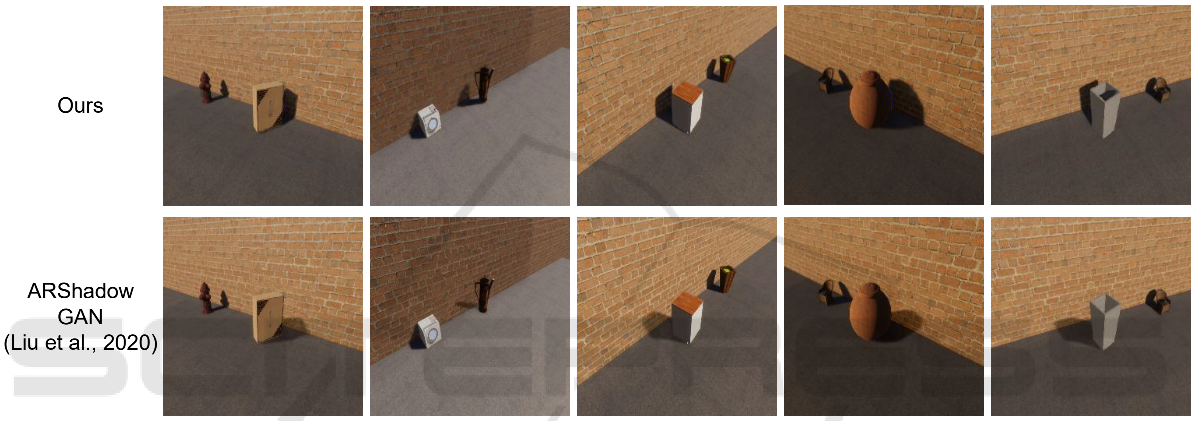
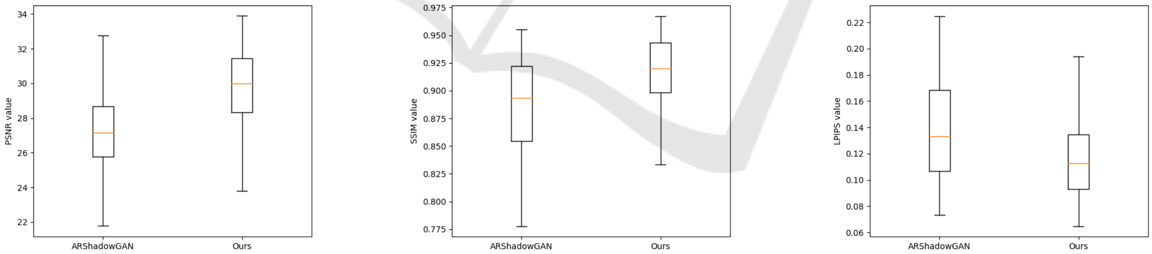


Figure 5: A comparison of our method and ARShadowGAN.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS



Experiment PSNR

Experiment SSIM

Experiment LPIPS

Figure 6: Box plots summarizing the results for each case in the experiment.

5 EXPERIMENT

For the learning process, 3,600 images in the dataset were split into training, validation, and test sets in a ratio of 2880:360:360, respectively.

Figure 4 shows examples of the shadow generation results with the ground truth images. Also, Figure 5 shows comparative images with ARShadowGAN (Liu et al., 2020). The quantitative evaluation using PSNR, SSIM, and LPIPS (Zhang et al., 2018) for ARShadowGAN is displayed in box plots in Figure 6. The numerical comparison is presented in Table 1.

6 ABLATION STUDY

An ablation study was conducted to examine the impact of each loss term by removing L_{per} and L_{adv} respectively. Figure 7 shows The ablation study results.

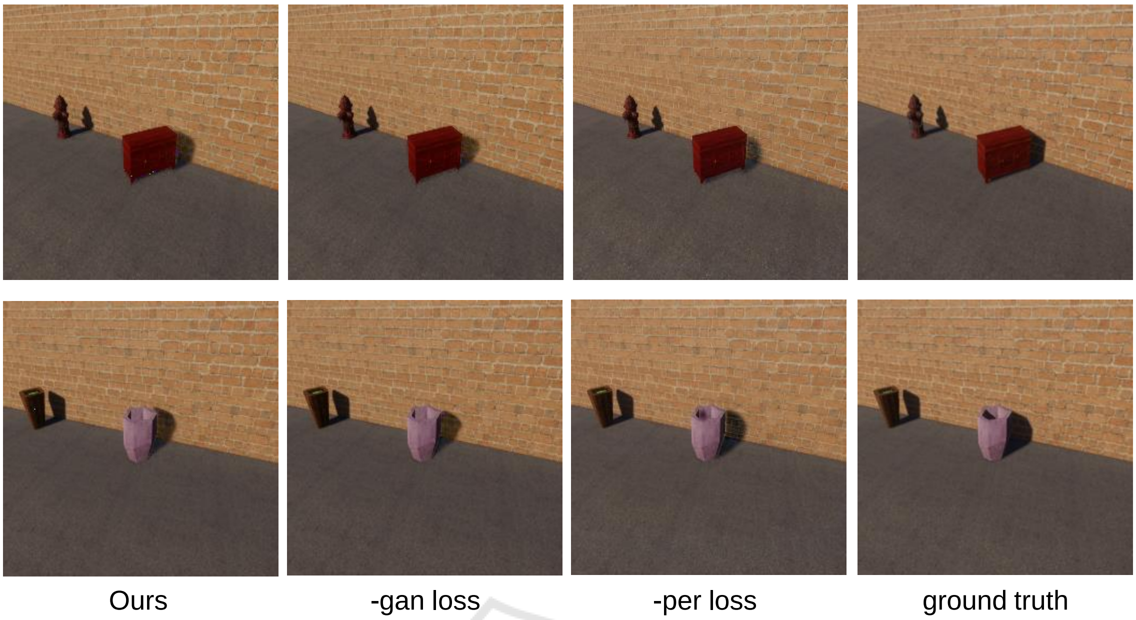
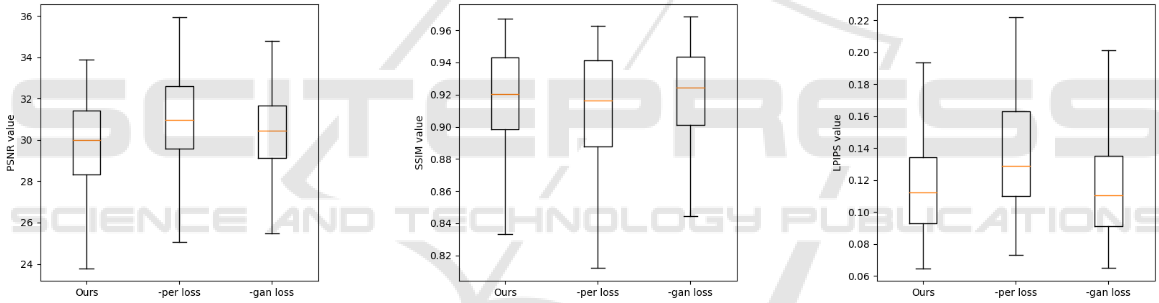


Figure 7: A comparison between the regular results in the ablation study, the results with the perceptual loss removed, the results with the GAN loss removed, and the ground truth images.



Ablation Study PSNR

Ablation Study SSIM

Ablation Study LPIPS

Figure 8: Box plots summarizing the results for each case in the ablation study.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 1: Experiment Average Rating.

	PSNR [dB]	SSIM	LPIPS
ARShadowGAN	28.98 ± 2.29	0.9021 ± 0.0436	0.05544 ± 0.0210
Ours	29.57 ± 2.53	0.9176 ± 0.0295	0.1149 ± 0.0268

Table 2: Ablation Study Average Rating.

	PSNR [dB]	SSIM	LPIPS
Ours	29.57 ± 2.53	0.9176 ± 0.0295	0.1149 ± 0.0268
-per loss	30.96 ± 2.34	0.9115 ± 0.0343	0.1358 ± 0.0261
-gan loss	30.33 ± 2.08	0.9199 ± 0.0286	0.1138 ± 0.0289

Figure 8 shows the comparison of PSNR, SSIM, and LPIPS (Zhang et al., 2018) box plots. Table 2 shows the average and standard deviation in each condition.

7 DISCUSSION

Figure 4 shows shadow generation results by our method. It was observed that for the virtual object of a chair, even the legs of the chair were projected well as a realistic shadow, indicating that the method were able to handle relatively detailed objects. Figure 5 shows a comparison between our method and ARShadowGAN. The figure indicates our method is capable to generate shadows with non-planar backgrounds. However, we can also see points to be improved. In the output images of our method, there were issues such as the failure to generate shadows for the convex parts of objects, excessive noise when projecting shadows onto walls, and shadow overlapping on concave parts of objects, which tended to cause noise. Specifically, when projecting shadows onto walls, noise was more prevalent for areas with lower brightness. This is likely because the color of the wall and the projected shadow were similar, leading to poor learning in the Perception Loss. Additionally, the issue of noise in shadows within the object is thought to be caused by the absence of the virtual object's image in the depth map. Since the mask image of the virtual object was provided, the mask region in the depth map became unnecessary information, possibly leading to poor learning in those regions.

For the qualitative assessment of the ablation study, when comparing the images in Figure 7, it was confirmed that without the perceptual loss, the shadows around the object's outline were not generated compared to the results of our method, resulting in unnatural images. Without the GAN loss, the shadows projected onto the wall were shallower in angle and smaller in size. For the quantitative assessment of the results in Table 2 and Figure 8, the data did not follow a normal distribution. It was confirmed by the Shapiro-Wilk test. When performing the Mann-Whitney U test between Ours and the -per loss condition, a significant difference was found in PSNR and SSIM at $p = 0.05$. This suggests that the perceptual loss term contributes to noise reduction and structural similarity in the images. Furthermore, a significant difference was found only in PSNR at $p = 0.05$ between Ours and -gan loss, indicating that the Discriminator loss term likely contributes to noise reduction.

8 LIMITATION

One limitation of this study is that the dataset only includes vertical walls, so accuracy may decrease depending on the complexity of the backgrounds. Additionally, since the shape information beyond the outline of the virtual objects is not included, generating shadows for complex virtual objects remains difficult. To address this challenge, it will be necessary to incorporate shape information of virtual objects in the learning process.

Moreover, the acquisition of depth images from real-world environments is still imprecise, which means that the model used in this study may not achieve sufficient accuracy when applied in real-world scenarios. As a potential solution for applying this model in the real world, segmentation and labeling of the ground and walls could be used as an alternative input in place of depth images.

9 CONCLUSION

In this study, we constructed an MR dataset that includes depth images and generated shadows for virtual objects on non-planar background geometries. For the dataset construction, a new method for incorporating depth images as input was established. By utilizing the depth images, this study proved that it is possible to cast shadows of virtual objects onto surfaces beyond flat background. The results of this study suggest that generating shadows in consideration of depth information can be applied for complex background geometries.

REFERENCES

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chrysanthakopoulou, A. and Moustakas, K. (2024). Real-time shader-based shadow and occlusion rendering in ar. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 969–970.

Couturier, A. (2023). Scenecity: 3d city generator addon for blender. https://www.cgchan.com/store/scenecity. 5.5.2024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hoffman, H., Hollander, A., Schroder, K., et al. (1998). Physically touching and tasting virtual objects enhances the realism of virtual experiences. *Virtual Reality*, 3:226–234.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham. Springer International Publishing.

Liu, D., Long, C., Zhang, H., Yu, H., Dong, X., and Xiao, C. (2020). Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8136–8145.

Meng, Q., Zhang, S., Li, Z., Wang, C., Zhang, W., and Huang, Q. (2023). Automatic shadow generation via exposure fusion. *IEEE Transactions on Multimedia*, 25:9044–9056.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241, Cham. Springer.

Schrater, P. and Kersten, D. (2000). How optimal depth cue integration depends on the task. *International Journal of Computer Vision*, 40:71–89.

Sheng, Y., Zhang, J., Philip, J., Hold-Geoffroy, Y., Sun, X., Zhang, H., Ling, L., and Benes, B. (2023). Pixht-lab: Pixel height based light effect generation for image compositing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16643–16653.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS