Joint Inpainting of RGB and Depth Images by Generative Adversarial Network with a Late Fusion approach

Ryo Fujii* Keio University
Ryo Hachiuma† Keio University
Hideo Saito‡ Keio University

Abstract

Image inpainting aims to restore texture of missing regions in scene from an RGB image. In this paper, we aim to restore not only the texture but also the geometry of the missing regions in scene from a pair of RGB and depth images. Inspired by the recent development of generative adversarial network, we employ an encoder–decoder-based generative adversarial network with the input of RGB and depth image. The experimental results show that our method restores the missing region of both RGB and depth image.

Index Terms: Artificial intelligence—Computer vision—Computer vision tasks—Scene understanding; Computer graphics—Graphics systems and interfaces—Mixed / augmented reality

1 Introduction

Diminished reality aims to remove real objects from images and fill in the removed regions with plausible textures. Removed regions are restored from multi-view observations [7], or inpainting using pixels in the image [5]. As the multi-view-based methods utilize the pixels directly observed from another camera, they can provide accurate restoration. However, multi-view based methods cannot restore unobserved areas. On the other hand, the inpainting-based methods use the pixels in the image to fill in the pixels of the removed regions, so that they do not need other cameras and recorded observation. If the filled pixels are plausible, the inpainting-based methods have an advantage over the multi-view based methods.

Image inpainting is the task to restore the missing region in an image with plausible contents based on the surrounding context. This allows restoration of damaged images or removal of unwanted objects or occluded regions from the RGB image. Inpainting of RGB image can restore the missing region texture in the scene, but it cannot handle the geometric structure restoration. In this paper, we aim to achieve the restoration of not only the missing region texture but also the geometry of it.

Traditional image inpainting methods represent diffusion-based approaches or patch-based approaches that make use of low-level features. The diffusion-based methods [1] propagate the texture from the surrounding context to the region of interest. However, diffusion-based approaches can only fill small or narrow holes. The patch-based methods, such as PatchMatch [2], can deal with larger and more complicated image completion. Mori *et al*. have presented 3D PixMix [8], which addresses non-planar scenes by using both color and depth information in the inpainting process. However, they are unable to restore objects which are not found in the image.

Rapid development of deep convolution neural networks and deep generative models inspired recent image inpainting tasks. In particular, the generative adversarial network (GAN) [3] based methods have been used for recent image inpainting efforts. Iizuka *et al*. [4] improved the consistency of the image inpainting result by introducing global and local discriminator. In this paper, we adopt global and local discriminator and the residual blocks with dilation factor. The residual blocks allow the structure to go deeper to improve performance without facing degradation or vanishing gradients.

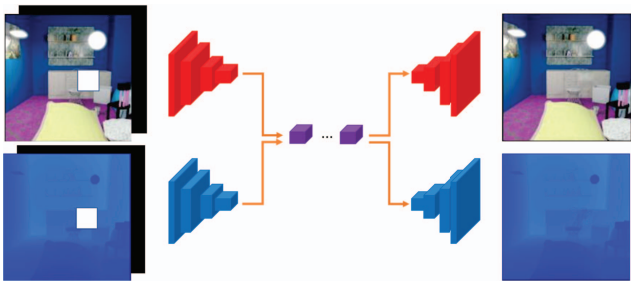

Figure 1: Illustration of completion network. The network takes the input of RGB and depth images with missing regions, and it jointly outputs restored RGB and depth images.

One simple solution to restore both texture and geometry of the missing regions is to train the two networks independently; the one restores textures with the input of RGB image and the other restores geometry with the input of depth image. Inspired by the recent object recognition method [9], we aim to construct an inpainting network that exploits the complementary relationship between RGB and depth information for RGB-D inpainting. Wang *et al*. [9] showed that the late fusion approach, which combines the extracted features of RGB and depth image, improves the classification accuracy of the objects in the images. Therefore, we also employ the late fusion approach to use RGB and depth information.

We present a method to jointly restore texture and geometry of the scene. Our method is based on GAN with the input of a pairs of RGB and depth images. We employ late fusion approach to fuse RGB and depth information. The experimental results show that our method successfully restored the missing regions of both RGB and depth image.

Our contributions are as follows:we first propose a deep learning architecture that jointly restores the texture and geometry of the scene from RGB and depth image. Second, we employ late fusion approach to fuse RGB and depth information, which exploits the complementary relationship between RGB and depth image.

2 Method

Our goal is to fill in missing regions of both RGB and depth images with plausible textures or geometries. We propose the GAN-based late fusion architecture to utilize each feature as complementary information. Our network consists of two parts, a completion network and a discriminator network.

Completion network. Figure 1 shows the completion network architecture. Our completion network consists of an RGB encoder-decoder, a depth encoder-decoder and a fusion part. RGB and depth encoders consist of four down-sampling convolution layers and both decoders consist of four up-sampling convolution layers. The kernel

*e-mail: ryo.fujii0112@keio.jp
†e-mail: ryo-hachiuma@hvrl.ics.keio.ac.jp
‡e-mail: hs@keio.jp

978-1-7281-4765-9/19/$31.00 ©2019 IEEE
DOI 10.1109/ISMAR-Adjunct.2019.00063

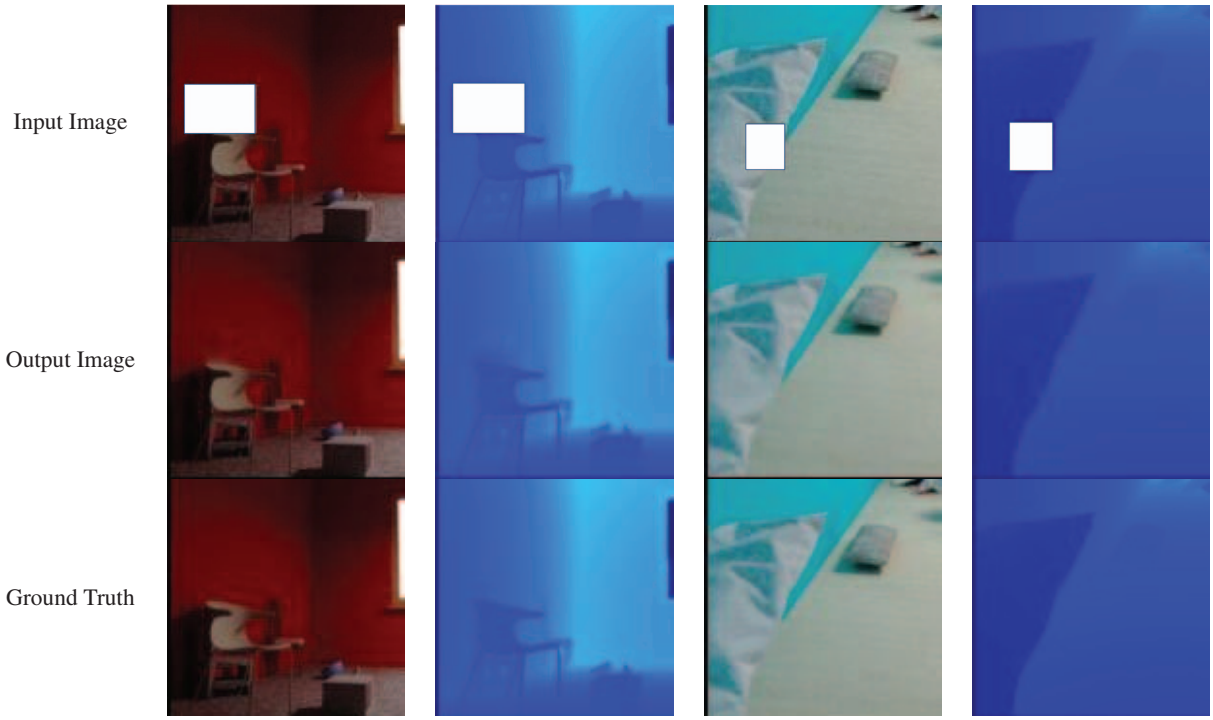

Figure 2: The results of the proposed inpainting network. Missing regions in the images of the first row are colored in white. We cropped images to visualize them clearly.

sizes are seven for the first and last layers and three for the remaining layers. We use sigmoid function after the last decoder layers and ReLU and batch normalization after the remaining convolutional layers. Added features of each encoder are used as fusion part input. The fusion part consists of nine residual blocks with dilation factors. We set the dilation factor of the first three blocks at two and duplicate it by three blocks.

Discriminator network. We employ a discriminator network proposed by Iizuka *et al*. [4]. It consists of two networks a local discriminator and a global discriminator. The global discriminator judges the scene consistency and the local discriminator assesses the quality of small completed area. The discriminators input is four channels, RGB-D data.

Training. To optimize our network, we combine mean squared error (MSE) loss and generative adversarial loss [3] as proposed by Iizuka *et al*. [4]. We add RGB and depth MSE loss. Moreover, the discriminator loss is calculated with four-channel input. We combined these two loss functions to optimize the whole network.

3 Experiment

We evaluate our network with SceneNet RGB-D dataset [6], which consists of 5M rendered RGB-D images from over 15K trajectories in synthetic layouts. The pose of the objects is randomly arranged and physically simulated with random lighting, camera trajectories, and textures. We train our model using about 2M images taken from the SceneNet RGB-D. First, we trained the inpainting network for $33,750$ iterations, and then we trained the discriminator for $3,750$ iterations. Finally, we jointly trained our model for $150,000$ iterations. In Figure 2, we show the inpainting result of two scene from SceneNet RGB-D. We confirmed that our network restores the missing region of both RGB and depth image. This shows that the late fusion approach made the edge of each restored region clearly.

4 Conclusion

In this work, we proposed a GAN-based RGB-D encoder–decoder inpainting network. This inpainting employs late fusion architecture that consists of residual blocks with dilation factors. Our network jointly restores the missing region of RGB and depth image. In the future work, we will conduct quantitative comparison with independent networks and other methods.

References

[1] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera. Filling-in by joint interpolation of vector fields and gray levels. *Trans. Img. Proc.*, 10(8):1200–1211, Aug. 2001.

[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *SIGGRAPH*, 28(3), Aug. 2009.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds., *NIPS*, pp. 2672–2680. 2014.

[4] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and Locally Consistent Image Completion. *SIGGRAPH*, 36(4):107:1–107:14, 2017.

[5] N. Kawai, T. Sato, and N. Yokoya. Diminished reality based on image inpainting considering background geometry. *IEEE Trans. Vis. Comput. Graphics*, 22:1–1, Jan. 2015.

[6] J. McCormac, A. Handa, S. Leutenegger, and A. J.Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *ICCV*, 2017.

[7] S. Meerits and H. Saito. Real-time diminished reality for dynamic scenes. In *Proc. of the IEEE ISMARW*, pp. 53–59, 2015.

[8] S. Mori, J. Herling, W. Broll, N. Kawai, H. Saito, D. Schmalstieg, and D. Kalkofen. 3d pixmix: Image-inpainting in 3d environments. In *Adjunct Proc. of the IEEE ISMAR*, 2018.

[9] A. Wang, J. Lu, J. Cai, T. Cham, and G. Wang. Large-margin multi-modal deep learning for rgb-d object recognition. *IEEE Trans. Multimedia*, 17(11):1887–1898, Nov 2015.