Layout-guided Indoor Panorama Inpainting with Plane-aware Normalization

Chao-Chen Gao[1][0000−0002−5325−3592], Cheng-Hsiu Chen[1][0000−0002−8652−5830], Jheng-Wei Su[1][0000−0003−3148−002X], and Hung-Kuo Chu[1][0000−0001−7153−4411]

[1]National Tsing Hua University, Taiwan
hkchu@cs.nthu.edu.tw

(a) Synthetic empty scene (b) Synthetic furnished scene (c) Real-world empty scene (d) Real-world furnished scene

Fig. 1: **Indoor panorama inpainting.** We present a learning-based indoor panorama inpainting method that is capable of generating plausible results for the tasks of hole filling (a)(c) and furniture removal (b)(d) in both synthetic (a)(b) and real-world (c)(d) scenes.

Abstract. We present an end-to-end deep learning framework for indoor panoramic image inpainting. Although previous inpainting methods have shown impressive performance on natural perspective images, most fail to handle panoramic images, particularly indoor scenes, which usually contain complex structure and texture content. To achieve better inpainting quality, we propose to exploit both the global and local context of indoor panorama during the inpainting process. Specifically, we take the low-level layout edges estimated from the input panorama as a prior to guide the inpainting model for recovering the global indoor structure. A plane-aware normalization module is employed to embed plane-wise style features derived from the layout into the generator, encouraging local texture restoration from adjacent room structures (i.e., ceiling, floor, and walls). Experimental results show that our work outperforms the current state-of-the-art methods on a public panoramic dataset in both qualitative and quantitative evaluations. Our code is available online[1].

---

[1] https://ericsujw.github.io/LGPN-net/

1   Introduction

Image inpainting is a widely investigated topic in computer graphics and vision communities, which aims at filling in missing regions of an image with photorealistic and fine detailed content. It plays a crucial step toward many practical applications, such as image restoration, object removal, etc. With the rapid development of deep learning, image inpainting has been revisited and improved significantly in the past few years. A considerable body of researches has been explored to generate impressive results on perspective datasets.

In this work, we address the image inpainting problem in the context of indoor panoramas. Indoor panoramas provide excellent media for the holistic scene understanding [40] that would further benefit several applications such as object detection, depth estimation, furniture rearrangement, etc. In particular, removing foreground objects and filling the missing regions in an indoor panorama is essential for the interior redesign task. However, the complex structures and textures presented in the indoor scenes make the inpainting problem non-trivial and challenging for previous methods. As shown in Figure 2(EC), results generated by a state-of-the-art deep learning method fail to align the image structure along the layout boundaries and produce inconsistent blurry image contents.

Recently, Gkitsas et al. [9] introduced PanoDR, a diminished reality-oriented inpainting model for indoor panorama. The main idea is to translate a furnished indoor panorama into its empty counterpart via a network that leverages both a generator and an image-to-image translation module. The inpainting result is then obtained by compositing the predicted empty panorama and input panorama using the object mask. However, there are still obvious artifacts near the boundaries of masked regions as shown in Figure 2.

To achieve better inpainting quality, we present an end-to-end deep generative adversarial framework that exploits both the global and local context of indoor panoramas to guide the inpainting process. Specifically, we take the low-level layout boundaries estimated from input panorama as a conditional input to guide the inpainting model, encouraging the preservation of sharp boundaries in the filled image. A plane-aware normalization module is then employed to embed local plane-wise style features derived from the layout into the image decoder, encouraging local texture restoration from adjacent room structures (i.e., ceiling, floor, and individual walls). We train and evaluate our model on a public indoor panorama dataset, Structured3D [41]. Experimental results show that our method produces results superior to several state-of-the-art methods (see Figure 1, Figure 2 and Figure 5). The main contributions are summarized as follows:

- We present an end-to-end generative adversarial network that incorporates both the global and local context of indoor panoramas to guide the inpainting process.
- We introduce a plane-aware normalization module that guides the image decoder with spatially varying normalization parameters per structural plane (i.e., ceiling, floor, and individual walls).
- Our method achieves state-of-the-art performance and visual quality on synthetic and real-world datasets.
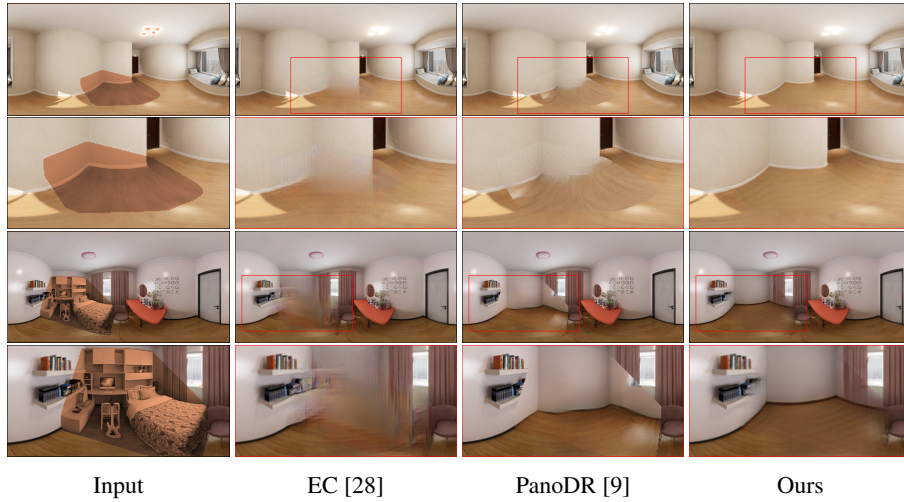
Fig. 2: **Limitations of existing methods.** EC [28] and PanoDR [9] fail to align the image structure along the layout boundaries and produce inconsistent blurry image contents in the inpainted regions (red mask).

2 Related Work

Traditional image inpainting. There are two main genres among traditional image inpainting works: diffusion-based methods and patch-based methods. Diffusion-based methods [6,1,2,4,23,36] propagate pixels from neighboring regions to the missing ones to synthesize the image content. On the other hand, patch-based methods [3,32,19,11,27,26,7] fill the missing regions by searching for and copying similar image patches from the rest of the image or existing image datasets. Without a high-level understanding of the image contents, these methods easily fail on images with complex structures.

Learning-based image inpainting. With the rapid development of deep learning, several image inpainting techniques based on convolutional neural networks (CNN) have been proposed. These methods aim to learn a generator from a large dataset to produce photorealistic image contents in the missing regions effectively. Context Encoders [30] pioneers CNN-based image inpainting by proposing an adversarial network with an encoder-decoder architecture. However, due to the information bottleneck layer of the autoencoder, the results are often blurry, incoherent, and can not work on irregular masks. Yu et al. [39] proposed a coarse-to-fine network and a context-aware mechanism to reduce blurriness. Iizuka et al. [14] adopted local and global discriminators and used dilated convolutions to increase the model's receptive field and enhance coherence. Liu et al. [25] proposed partial convolutions, which only consider valid pixels during convolution, to handle irregular masks. Yu et al. [38] further extends the partial convolutions by introducing a dynamic feature gating mechanism, named gated convolutions, to deal with free-from masks. Both Liu et al. [25] and Yu et al. [38] adopt PatchGAN discriminator [15] to improve the coherence further. Recently, several models were pro-

posed to significantly improve the image painting quality by incorporating the structure knowledge in a different context, including image edges [28,22], object contours [37], smooth edge-preserving map [31], and gradient map [17]. Nazeri et al. [28] introduced a two-stage network named EdgeConnect, which firstly recovers the missing edges in the masked regions, followed by a generator conditioned on the reconstructed edge map. The authors prove that the structure-to-content approach can effectively preserve the structure information in the inpainting results. However, EdgeConnect uses canny edges to represent structure features, which might be suitable for natural images but may lead to complex local edges in indoor scenes. In contrast, our work exploits the Horizon-Net [34] to estimate layout edges, representing the global room structure, which is suitable for our indoor inpainting task. In addition, our model is an end-to-end architecture instead of a two-stage network. Yang et al. [17] developed a multi-task learning framework to jointly learn the completion of image contents and structure map (edges and gradient). A structure embedding scheme is employed to embed the learned structure features while inpainting explicitly. The model further learns to exploit the recurrent structures and contents via an attention mechanism. While demonstrating impressive performance in generating realistic results, these structure-aware methods still fail to model long-range structure correspondence such as the layout in the indoor scenes. On the other hand, some works have successfully recovered a single partially occluded object [5,20]. However, their architecture does not handle multiple object instances of the same class and is thus not suitable for our context where the plane-wise segmentation consists of different numbers of wall planes.

Image-to-image translation. The image inpainting is essentially a constrained image-to-image translation problem. Significant efforts have been made to tackle various problems based on image-to-image translation architectures [15,42,18]. Here we focus on the ones that are closely related to our work. Park et al. [29] introduced SPADE, which utilizes a spatial adaptive normalization layer for synthesizing photorealistic images given an input semantic segmentation map. Specifically, a spatially-adaptive learned transform modulates the activation layer with a semantic segmentation map and effectively propagates the semantic information throughout the network. In contrast to SPADE, which uses only one style code to control the image synthesis, Zhu et al. [43] presents SEAN by extending the SPADE architecture with per-region style encoding. By embedding one style code for individual semantic classes, SEAN shows significant improvement over SPADE and generates the highest quality results. In the context of indoor scenes, Gkitsas et al. [9] introduce PanoDR that combines image-to-image translation with a generator to constrain the image inpainting with the underlying scene structure. Percisely, to convert a furnished indoor panorama into its empty counterpart, PanoDR exploits a generator for synthesizing photorealistic image contents where the global layout structure is preserved via an image-to-image translation module. The empty indoor panorama is then used to complete the masked regions in the input panorama via a simple copy-and-paste process. Gkitsas et al. [10] extend the architecture of PanoDR to make the model end-to-end trainable. However, the quantitative evaluation indicates that the performance improvement is marginal compared with PanoDR. Our system also combines a generator with image-to-image translation as PanoDR does. However, we obtain superior results than PanoDR by exploiting the
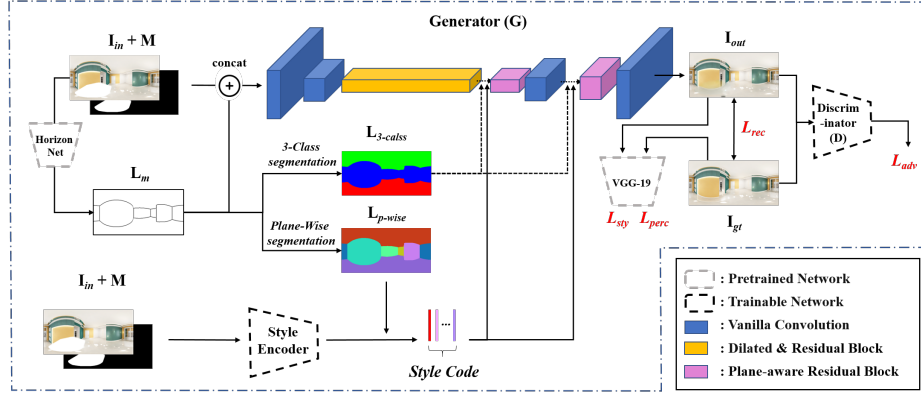
Fig. 3: **Architecture overview.** Our network architecture follows the conventional generative adversarial network with an encoder-decoder scheme supervised by low- and high-level loss functions and a discriminator. Given a masked indoor panoramic image $\mathbf{I}_{in}$ with a corresponding mask $\mathbf{M}$, our system uses an off-the-shelf layout prediction network to predicts a layout map. The low-level boundary lines in $\mathbf{L}_m$ serve as a conditional input to our network to assist the inpainting. Then, we compute two semantic segmentation maps from the layout map $\mathbf{L}_m$, declared $\mathbf{L}_{3-class}$ and $\mathbf{L}_{p-wise}$, where the latter is used to generate plane-wise style codes for ceiling, floor, and individual walls. Finally, these per plane style codes, together with $\mathbf{L}_{3-class}$, are fed to a structural plane-aware normalization module to constrain the inpainting.

global layout edges as a prior and adapting SEAN blocks in a local plane-wise manner to guide the inpainting. Moreover, in contrast to PanoDR performs the inpainting task via an indirect way, our system performs the inpainting task in an end-to-end fashion, directly completing the mask areas instead of hallucinating an empty scene, thus resulting in better visual quality and consistency.

3 Overview

Figure 3 illustrates an overview of our architecture. Our system takes a masked panoramic image $\mathbf{I}_{in}$ and the corresponding binary mask $\mathbf{M}$ as inputs and generates the inpainted panoramic image $\mathbf{I}_{out}$. The masked panoramic image is generated by $\mathbf{I}_{in} = \mathbf{I}_{gt} \odot (\mathbf{1} - \mathbf{M})$, where $\mathbf{I}_{gt}$ represents the ground-truth panoramic image and $\odot$ denotes the Hadamard product. Our system first utilizes an off-the-shelf model to estimate the room layout $\mathbf{L}_m$ from input masked panoramic image. This layout map is then concatenated with $\mathbf{I}_{in}$ and $\mathbf{M}$ to obtain a five-channel input map fed into the generator $\mathbf{G}$. We further derive two semantic segmentation maps $\mathbf{L}_{3-class}$ and $\mathbf{L}_{p-wise}$ using the layout map for the subsequent normalization module (Section 4.1). The image generation model follows the conventional generative adversarial architecture with one content encoder and one image decoder with one discriminator. (Section 4.2). To impose structure information during inpainting, we introduce a plane-aware normalization that modifies
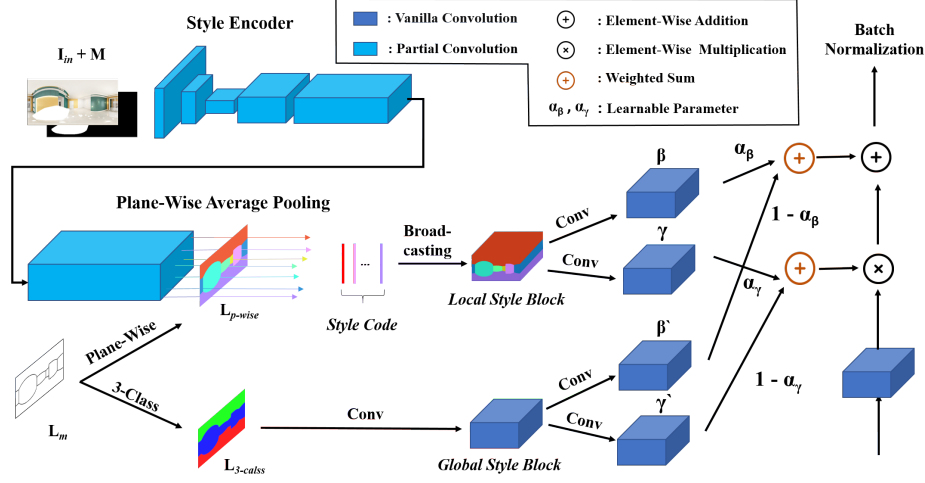
Fig. 4: **Plane-aware normalization.** Given an incomplete indoor panoramic image $\mathbf{I}_{in}$ with mask $\mathbf{M}$, we first predict two normalization values β and γ through several partial convolution [25] blocks and a plane-wise average pooling based on the plane-wise segmentation map $\mathbf{L}_{p-wise}$. Second, we predict another set of normalization values β' and γ' through several vanilla convolution blocks based on the 3-class segmentation map $\mathbf{L}_{3-class}$. The final normalization values are thus computed using the weighted sum weighted by learnable parameters α_β and α_γ.

the SEAN [43] block with two semantic segmentation maps to guide the decoder with spatially varying normalization parameters per structural plane (i.e., ceiling, floor, and individual walls). Such a plane-aware normalization provides useful guidance for global structure preservation as well as consistent local image content generation (Section 4.3). Finally, common loss functions in image inpainting, including the reconstruction loss, the perceptual loss, the style loss, and the adversarial loss are employed to train our model (Section 4.4).

4   Method

4.1   Layout Guidance Map

We employ an off-the-shelf model, HorizonNet [34], to estimate a layout map from input masked panorama. Through a recurrent neural network, the HorizonNet predicts a 3-dimensional vector representing ceiling, ground, and corner location. We further process the output vector to generate a layout map $\mathbf{L}_m$ comprising low-level boundary lines. This layout map serves as a conditional input to encourage the preservation of global layout structure while inpainting. Moreover, we extract two semantic segmentation maps from the layout map that depict (i) the segmentation mask $\mathbf{L}_{3-class}$ with three semantic labels of indoor scene, i.e., ceiling, floor, and wall; and (ii) a plane-wise segmentation mask $\mathbf{L}_{p-wise}$ where pixels are indexed in a per structural plane basis

(i.e., ceiling, floor, or individual walls). These semantic segmentation maps are generated using conventional image processing operations (i.e, flood-fill) and will be used in the later normalization module.

4.2   Image Inpainting Backbone

As shown in Figure 3, our network architecture consists of one generator and one discriminator. The generator $\mathbf{G}$ follows a conventional scheme with one content encoder and one image decoder. The content encoder consists of two down-sampling convolution blocks followed by eight residual blocks using dilated convolution [12]. The image decoder uses a cascade of our proposed plane-aware residual blocks and two up-sampling blocks. Motivated by EdgeConnect [28], we use PatchGAN [16] as our discriminator to determine the real or fake sample by dividing the input image into several patches. In the following sections, we will elaborate plane-aware residual block, loss functions, and discriminator in more detail.

4.3   Plane-aware Normalization

Considering the different styles among wall planes is very common in real-world indoor scenes. We follow the architecture of SEAN [43] and propose leveraging two kinds of segmentation maps $\mathbf{L}_{p-wise}$ and $\mathbf{L}_{3-class}$ to establish our plane-aware normalization (see Figure 4). Our plane-aware normalization consists of one style encoder and two style blocks, which enhance the global style semantics and local style consistency of the generated results. The inputs of the style encoder include masked panoramic image $\mathbf{I}_{in}$ and mask image $\mathbf{M}$. We use partial convolution blocks in style encoder instead of vanilla convolution to make feature extraction conditioned only valid pixels. We first adopt the plane-wise average pooling on the output features to generate style codes for each plane based on $\mathbf{L}_{p-wise}$. Second, we spatially broadcast each style code on the corresponding area and output the local style block. On the other side, we predict the global style block by passing the 3-class segmentation map $\mathbf{L}_{3-class}$ through several convolution layers. Finally, the remaining part of our plane-aware normalization follows the same architecture of SEAN [43], and combines global and local style blocks into the downstream β and γ parameters of the final batch normalization.

4.4   Loss Functions

Here we elaborate on the low- and high-level loss functions and the discrimination used for training our image generator.

Reconstruction loss measures the low-level pixel-based loss between the predicted and ground-truth images. To encourage the generator to pay more attention to the missing regions, we additionally calculate the L_1 loss in the missing regions. The reconstruction loss L_{rec} is defined as follows:

$$L_{rec} = \|\mathbf{M} \odot \mathbf{I}_{gt} - \mathbf{M} \odot \mathbf{I}_{out}\|_1 + \|\mathbf{I}_{gt} - \mathbf{I}_{out}\|_1 , \tag{1}$$

where $\mathbf{I}_{gt}$ and $\mathbf{I}_{out}$ represent the ground-truth image and the generator's output, respectively, and $\mathbf{M}$ is a binary mask.

Perceptual loss encourages the predicted and ground-truth images to have similar representation in high-level feature space extracted via a pre-trained VGG-19 [33], and is defined as follows:

$$L_{perc} = \sum_i \left\| \phi_i \left(\mathbf{I}_{gt} \right) - \phi_i \left(\mathbf{I}_{out} \right) \right\|_1, \tag{2}$$

where ϕ_i is the activation map of the ith layer of the pre-trained feature extraction network.

Style loss calculates the co-variance difference between the activation maps. For the activation map ϕ_i of size $C_i \times H_i \times W_i$, the style loss is defined as follows:

$$L_{sty} = \left\| G_i^{\phi} \left(\mathbf{I}_{gt} \right) - G_i^{\phi} \left(\mathbf{I}_{out} \right) \right\|_1, \tag{3}$$

where G_i^{ϕ} is a $C_i \times C_i$ gram matrix [8] constructed by the activation map ϕ_i.

Adversarial loss is implemented with the patch-based discriminator [16], which outputs the feature map divided into several feature patches and uses hinge loss [24] to optimize the generator G and the discriminator D. The adversarial loss for generator G and discriminator D are defined as follows:

$$L_G = -D \left(\mathbf{I}_{out} \right), \tag{4}$$

$$L_D = \lambda_D \left(max \left(0, 1 + D \left(\mathbf{I}_{out} \right) \right) + max \left(0, 1 - D \left(\mathbf{I}_{gt} \right) \right) \right); \tag{5}$$

The overall loss function used in the generator G is defined as follows:

$$L_{total} = \lambda_{rec} L_{rec} + \lambda_{perc} L_{perc} + \lambda_{sty} L_{sty} + \lambda_G L_G, \tag{6}$$

where λ_{rec}, λ_{perc}, λ_{sty}, λ_G, and λ_D are the hyperparameters for weighting the loss functions.

5   Experiments

In this section, we evaluate the performance of our model by comparing it with several state-of-the-art image inpainting approaches and conducting ablation studies to verify the necessity of individual components in the proposed architecture. Please refer to our online webpage for other experiments and more results[2].

5.1   Experimental Settings

Dataset and baselines. We compare our model with the following state-of-the-art structure-aware image inpainting models:

---

[2] https://ericsujw.github.io/LGPN-net/

Input / Layout	EC [28]	LISK [17]	PanoDR [9]	Ours	GT

Fig. 5: **Qualitative comparisons with state-of-the-arts.** Top 8 rows: the inpainting results of the empty indoor scenes. Bottom 8 rows: the inpainting results of the furnished indoor scenes. Our method produces superior results in generating image contents that align the layout structure well and are consistent with the surrounding of the masked regions.

- EC [28]: a two-stage adversarial network that comprises an edge completion model followed by a generator.
- LISK [17]: a multi-task learning framework that exploits image structure embedding and an attention mechanism in the generator.
- PanoDR [9]: a deep learning framework that combines image-to-image translation with generator to condition the inpainting on the indoor scene structure.

The experiments were conducted on a public indoor panorama dataset, Structured3D [41], which contains 21,835 indoor panoramas. The official data split is adopted for training(18,362), validation(1,776), and testing(1,697). We follow the same procedure as PanoDR to generate mask images using contours of foreground furniture (see Section 3.1). We use the officially released implementation of baselines for training from scratch and testing. Note that each indoor panorama in Structured3D has two representations of the same scene (i.e., empty and furnished). Therefore, the experiments were conducted in two phases to evaluate our model and baselines in different application scenarios (i.e., structural inpainting vs. furniture removal).

Evaluation metrics. We take several commonly used image quality assessment metrics in previous inpainting approaches for quantitative evaluation. Specifically, we used the low-level feature metrics, including Mean Absolute Error (MAE), Peak Signal-to-Noise (PSNR), Structural Similarity Index (SSIM) [35], and Fréchet Inception Distance (FID) [13].

Implementation details. We implement our model in PyTorch and conduct the experiments on a single NVIDIA V100 with 32G VRAM. The resolution of the panoramic images is resized to 512×256. We use Adam [21] optimizer in the training process with the hyper-parameters setting of $b_1 = 0.0$ and $b_2 = 0.9$, a learning rate of 0.0001, and a batch size of 8. We empirically set $\lambda_{rec} = 1$, $\lambda_{perc} = 0.1$, $\lambda_{sty} = 250$, $\lambda_G = 0.1$, and $\lambda_D = 0.5$ in the total loss function (Equation 6). For HorizonNet [34], we use the official pre-trained model for layout estimation.

5.2   Evaluation on the Empty Scenes

In this experiment, we evaluate both the qualitative and quantitative performance of our model on the image inpainting task by comparing it with baselines. The qualitative comparisons are shown in Figure 5 (top 8 rows). In contrast to EC and LISK, which fail to restore image structures in the masked regions, our method faithfully generates image contents adhering to the underlying layout structure. While PanoDR shows slightly better structure preservation than EC and LISK, it fails to generate image contents consistent with the surrounding of masked regions as our method does. Therefore, our method achieves the best performance against all the baselines across all evaluation metrics as shown in Table 1 (top).

5.3   Evaluation on the Furnished Scenes

Furniture of irregular shape will more or less obscure the layout of the indoor scene, making it more challenging to restore the regular structure in the missing area. Therefore, in this experiment, we would like to evaluate how well our model learned from

Table 1: **Quantitative comparisons with state-of-the-arts.** The top and bottom tables summarize the performance of our model and baselines on the empty and furnished scenes, respectively.

Dataset	Method	PSNR↑	SSIM↑	MAE↓	FID↓
Empty scene	EC [28]	38.6936	0.9892	0.0039	3.9480
	LISK [17]	41.3761	0.9895	0.0055	4.1660
	PanoDR [9]	37.2431	0.9884	0.0040	4.3591
	Ours	**41.8444**	**0.9919**	**0.0030**	**2.5265**
Furnished scene	EC [28]	31.4439	0.9493	0.0076	11.9955
	LISK [17]	34.7325	0.9553	0.0068	14.2676
	PanoDR [9]	34.3340	0.9641	0.0051	7.8399
	Ours	**35.3923**	**0.9672**	**0.0047**	**7.2328**

Table 2: **Quantitative results of the ablation study.** We evaluate the effectiveness of our design choices by gradually adding the individual components into the architecture.

	PSNR ↑	SSIM ↑	MAE ↓	FID ↓
Backbone	40.6449	0.9911	0.0034	3.3915
Layout map only	41.2884	0.9916	0.0033	2.8105
Full model	**41.8444**	**0.9919**	**0.0030**	**2.5265**

empty scenes can generalize to the furnished scenes. Since the inpainting task setup here exactly matches the one defined in the PanoDR, we use the pre-trained model of PanoDR in this experiment for a fair comparison. As shown in Figure 5 (bottom 8 rows), our method still clearly outperforms baselines in generating image contents that align the layout structure well and are consistent with the surrounding of the masked regions. The quantitative results are shown in Table 1 (bottom). It is worth noting that the way PanoDR performs image completion via compositing the predicted empty image and input image using the object mask will lead to severe artifacts where occlusion occurred between foreground objects (see Figure 2(PanoDR)).

5.4   Ablation Study

Here, we conduct ablation studies to validate our model from different perspectives. First, we evaluate the necessity of individual design choices in our architecture. Then, we conduct two experiments to evaluate how sensitive our model is to the size of input masks and the quality of input layout maps.

Ablation on network architecture.   In this experiment, we start with the backbone model (**Backbone**) as the baseline, then progressively adding only layout guidance map (**Layout map only**), and our plane-aware normalization (**Full model**). As shown in Table 2, we obtain the best performance with the full model on all the metrics. The qualitative comparisons shown in Figure 6 indicate that adding layout guidance map generates clear structure boundaries in the final result (2nd and 3rd columns), while our

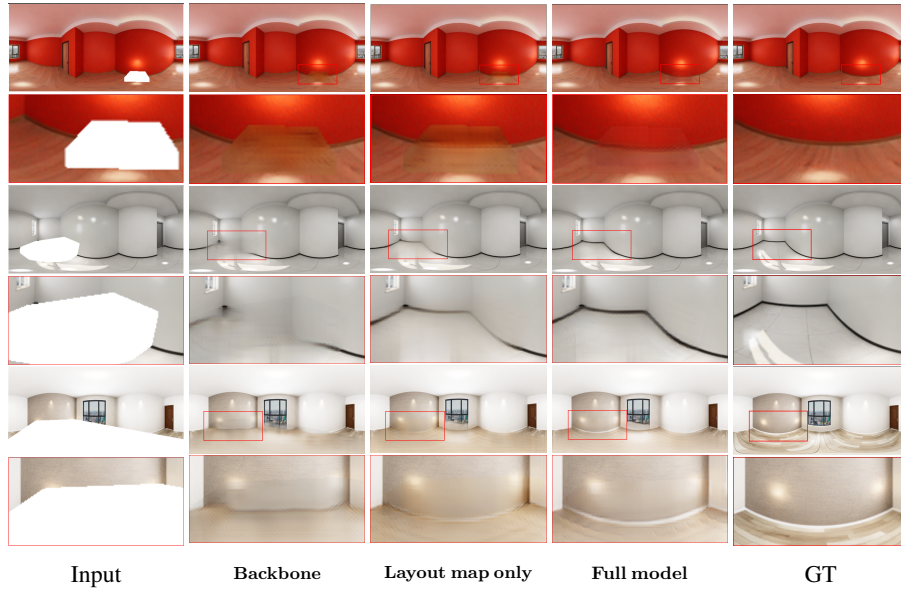| Input | Backbone | Layout map only | Full model | GT |

Fig. 6: **Qualitative results of the ablation study.** Side-by-side comparisons of inpainting results generated using our method by gradually adding individual components. From left to right, input images and masks, our baseline model (**Backbone**), adding the layout guidance map (**Layout map only**), full model with our plane-aware normalization (**Full model**), and ground truth images.

full model with plane-aware normalization can constrain the image generation to the adjacent structural planes and obtain visually consistent results (3rd and 4th columns).

Sensitivity to the mask size. In this experiment, we analyze the testing dataset and classify the images into different categories according to the area proportions of input masks. Table 3 shows the inpainting performance for each category. We can tell that the inpainting quality degrades with the increasing mask size. A significant drop occurs where the ratio of input mask is greater than 30%.

Sensitivity to the layout estimation. In order to explore the effect of the accuracy of layout estimation on the inpainting quality, we first devise a mechanism to generate layout maps with different levels of accuracy. Specifically, we feed masked images of different mask sizes into HorizonNet. We start by generating randomly located rectangle masks of 5% image size and increase the mask ratio to 10%, 30% and 50% to deliberately produce layout structures with decreasing quality. Then we take these layout maps as conditional inputs of our model and compare the inpainting performance empty-room testing dataset. As shown in Table 4, our model degrades marginally when the quality of estimated layouts decreases from 0.96 to 0.84, indicating our model is robust to the varying input layout maps.

Table 3: **Mask size vs. inpainting quality.**

Mask Size(%)	Count	Content		
		PSNR ↑	SSIM ↑	MAE ↓
0-10	1045	44.3921	0.9967	0.0011
10-20	163	34.2823	0.9841	0.0055
20-30	48	30.4371	0.9726	0.0111
30-40	39	25.0731	0.9386	0.0266
40+	13	24.2958	0.9345	0.0305
total	1308	41.8444	0.9919	0.0030

Table 4: **Accuracy of layout estimation vs. inpainting quality.**

Structure	Content			
mIOU ↑	PSNR ↑	SSIM ↑	MAE ↓	FID ↓
0.9603	42.3212	0.9925	0.0028	2.4322
0.9561	42.2871	0.9925	0.0028	2.4441
0.9175	42.0682	0.9923	0.0029	2.5624
0.8489	41.7300	0.9919	0.0030	2.8455

5.5 Qualitative Results on Real-world Scene

Real-world scenes have complex lighting and layout structure. However, the amount of data in the real-world scene dataset and the quality of furniture category annotations are insufficient for training our model, so we choose to train on the synthetic dataset Structured3D [41]. Nevertheless, we still compare our results with PanoDR [9], which also implements the furniture removal task, on the real-world scene dataset. Since the real-world scene dataset does not contain paired data (i.e., scenes before and after furniture removal), quantitative evaluation is infeasible and we can only provide qualitative comparisons here. Figure 7 shows that our inpainted results have a higher quality of structural maintenance and color restoration. Moreover, compared with PanoDR, we can still exert more stable performance in real-world scenes. Please refer to our online webpage for more results[3].

6   Conclusions

We proposed an end-to-end structural inpainting network for the indoor scene. We introduce layout boundary line conditions the output structure and utilize the plane-aware normalization to enhance planar style consistency. Experiment results show the outstanding performance of our model in both structural inpainting and furniture removal on the indoor scene.

---

[3] `https://ericsujw.github.io/LGPN-net/`

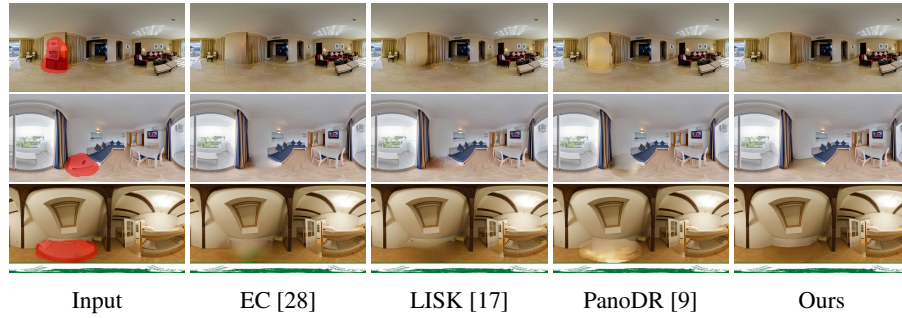| Input | EC [28] | LISK [17] | PanoDR [9] | Ours |

Fig. 7: **Qualitative comparisons with state-of-the-arts on real-world scenes.** Our model clearly outperforms baselines by preserving layout boundary and restoring local texture from adjacent room structures (i.e., floor and walls).
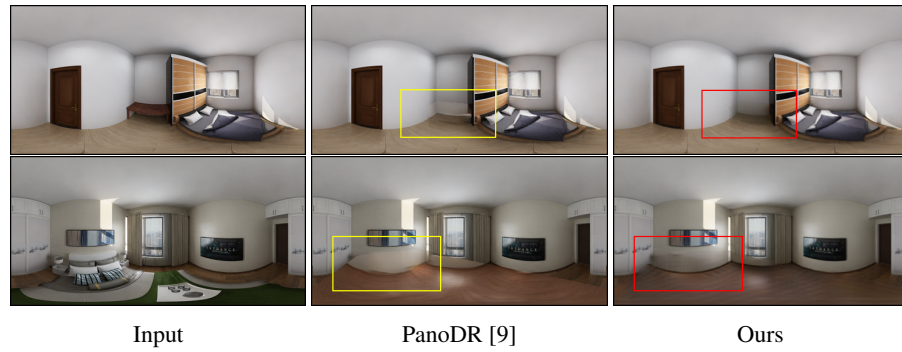

| Input | PanoDR [9] | Ours |

Fig. 8: **Limitation.** Both the state-of-the-art method and our model produce visual artifacts in the scenes presenting strong shading effect surrounding the removed furniture.

Limitations. In the real-world application of furniture removal, we can often see residuals of shading effect caused by the removed furniture. These residuals are hard to segment and even harder to model. As shown in Figure 8, our model is slightly affected by these residuals but still produces more realistic results than PanoDR [9].

Future work. We plan to adopt a more reasonable segmentation mask of the indoor scene inpainting which can cover the shading area and thus improve our results in those shaded scenes.

References

1. Ashikhmin, M.: Synthesizing natural textures. In: Proceedings of the 2001 Symposium on Interactive 3D Graphics. p. 217–226. I3D '01, Association for Computing Machin-

ery, New York, NY, USA (2001). https://doi.org/10.1145/364338.364405, `https://doi.org/10.1145/364338.364405`

2. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE Transactions on Image Processing **10**(8), 1200–1211 (2001). https://doi.org/10.1109/83.935036

3. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. In: ACM Transactions on graphics(TOG) (2009)

4. Drori, I., Cohen-Or, D., Yeshurun, Y.: Fragment-based image completion. ACM SIGGRAPH 2003 Papers (2003)

5. Ehsani, K., Mottaghi, R., Farhadi, A.: Segan: Segmenting and generating the invisible. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6144–6153 (2018). https://doi.org/10.1109/CVPR.2018.00643

6. Esedoglu, S.: Digital inpainting based on the mumford-shah-euler image model. European Journal of Applied Mathematics **13** (08 2003). https://doi.org/10.1017/S0956792502004904

7. Fan, Q., Zhang, L.: A novel patch matching algorithm for exemplar-based image inpainting. Multimedia Tools Appl. **77**(9), 10807–10821 (May 2018). https://doi.org/10.1007/s11042-017-5077-z, `https://doi.org/10.1007/s11042-017-5077-z`

8. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

9. Gkitsas, V., Sterzentsenko, V., Zioulis, N., Albanis, G., Zarpalas, D.: Panodr: Spherical panorama diminished reality for indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3716–3726 (2021)

10. Gkitsas, V., Zioulis, N., Sterzentsenko, V., Doumanoglou, A., Zarpalas, D.: Towards full-to-empty room generation with structure-aware feature encoding and soft semantic region-adaptive normalization. In: Farinella, G.M., Radeva, P., Bouatouch, K. (eds.) Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2022, Volume 4: VISAPP, Online Streaming, February 6-8, 2022. pp. 452–461. SCITEPRESS (2022). https://doi.org/10.5220/0010833100003124, `https://doi.org/10.5220/0010833100003124`

11. Guo, Q., Gao, S., Zhang, X., Yin, Y., Zhang, C.: Patch-based image inpainting via two-stage low rank approximation. IEEE Transactions on Visualization and Computer Graphics **24**(6), 2023–2036 (2018). https://doi.org/10.1109/TVCG.2017.2702738

12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018)

14. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and Locally Consistent Image Completion. ACM Transactions on Graphics (Proc. of SIGGRAPH) **36**(4), 107:1–107:14 (2017)

15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)

16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks (2018)

17. Jie Yang, Zhiquan Qi, Y.S.: Learning to incorporate structure knowledge for image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12605–12612 (2020)

18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019)

19. Kawai, N., Sato, T., Yokoya, N.: Diminished reality based on image inpainting considering background geometry. IEEE Transactions on Visualization and Computer Graphics **22**(3), 1236–1247 (2016). https://doi.org/10.1109/TVCG.2015.2462368

20. Ke, L., Tai, Y., Tang, C.: Occlusion-aware video object inpainting. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14448–14458. IEEE Computer Society, Los Alamitos, CA, USA (oct 2021). https://doi.org/10.1109/ICCV48922.2021.01420, `https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01420`

21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)

22. Li, J., He, F., Zhang, L., Du, B., Tao, D.: Progressive reconstruction of visual structure for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)

23. Liang, Z., Yang, G., Ding, X., Li, L.: An efficient forgery detection algorithm for object removal by exemplar-based image inpainting. J. Vis. Commun. Image Represent. **30**, 75–85 (2015)

24. Lim, J.H., Ye, J.C.: Geometric gan (2017)

25. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: The European Conference on Computer Vision (ECCV) (2018)

26. Liu, J., Yang, S., Fang, Y., Guo, Z.: Structure-guided image inpainting using homography transformation. IEEE Transactions on Multimedia **20**(12), 3252–3265 (2018). https://doi.org/10.1109/TMM.2018.2831636

27. Lu, H., Liu, Q., Zhang, M., Wang, Y., Deng, X.: Gradient-based low rank method and its application in image inpainting. Multimedia Tools Appl. **77**(5), 5969–5993 (Mar 2018). https://doi.org/10.1007/s11042-017-4509-0, `https://doi.org/10.1007/s11042-017-4509-0`

28. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning (2019)

29. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)

30. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: Feature learning by inpainting. In: Computer Vision and Pattern Recognition (CVPR) (2016)

31. Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G.: Structureflow: Image inpainting via structure-aware appearance flow (2019)

32. Ružić, T., Pižurica, A.: Context-aware patch-based image inpainting using markov random field modeling. IEEE Transactions on Image Processing **24**(1), 444–456 (2015). https://doi.org/10.1109/TIP.2014.2372479

33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)

34. Sun, C., Hsiao, C., Sun, M., Chen, H.: Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 1047–1056 (2019)

35. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861

36. Wei, Y., Liu, S.: Domain-based structure-aware image inpainting. Signal, Image and Video Processing **10** (07 2016). https://doi.org/10.1007/s11760-015-0840-y

37. Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., Luo, J.: Foreground-aware image inpainting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
38. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. arXiv preprint arXiv:1806.03589 (2018)
39. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. arXiv preprint arXiv:1801.07892 (2018)
40. Zhang, Y., Song, S., Tan, P., Xiao, J.: Panocontext: A whole-room 3d context model for panoramic scene understanding. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol. 8694, pp. 668–686. Springer (2014). https://doi.org/10.1007/978-3-319-10599-4_43
41. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: Proceedings of The European Conference on Computer Vision (ECCV) (2020)
42. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)
43. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)