BIPS: Bi-modal Indoor Panorama Synthesis via Residual Depth-aided Adversarial Learning

Changgyoon Oh[1]*, Wonjune Cho[2]*, Yujeong Chae[1]*, Daehee Park[1]*, Lin Wang[3], and Kuk-Jin Yoon[1]

[1] Visual Intelligence Lab., KAIST
{changgyoon, yujeong, bag2824, kjyoon}@kaist.ac.kr
[2] NAVER LABS
wonjune.cho@naverlabs.com
[3] AI Thrust, HKUST Guangzhou and Dept. of CSE, HKUST
linwang@ust.hk

Abstract. Providing omnidirectional depth along with RGB information is important for numerous applications. However, as omnidirectional RGB-D data is not always available, synthesizing RGB-D panorama data from limited information of a scene can be useful. Therefore, some prior works tried to synthesize RGB panorama images from perspective RGB images; however, they suffer from limited image quality and can not be directly extended for RGB-D panorama synthesis. In this paper, we study a new problem: RGB-D panorama synthesis under the various configurations of cameras and depth sensors. Accordingly, we propose a novel bi-modal (RGB-D) panorama synthesis (BIPS) framework. Especially, we focus on indoor environments where the RGB-D panorama can provide a complete 3D model for many applications. We design a generator that fuses the bi-modal information and train it via residual depth-aided adversarial learning (RDAL). RDAL allows to synthesize realistic indoor layout structures and interiors by jointly inferring RGB panorama, layout depth, and residual depth. In addition, as there is no tailored evaluation metric for RGB-D panorama synthesis, we propose a novel metric (FAED) to effectively evaluate its perceptual quality. Extensive experiments show that our method synthesizes high-quality indoor RGB-D panoramas and provides more realistic 3D indoor models than prior methods. Code is available at https://github.com/chang9711/BIPS.

Keywords: RGB-D panorama synthesis, indoor layout, GAN, VR/AR

1 Introduction

Omnidirectional RGB-D data is important for numerous applications, *e.g.*, VR/AR, yet it is not always available. Manually creating a 3D space from scratch is unrealistic and requires a huge effort, while capturing and restoring whole real-world requires high computational cost [35]. Synthesizing RGB-D panorama

* Equal contribution.

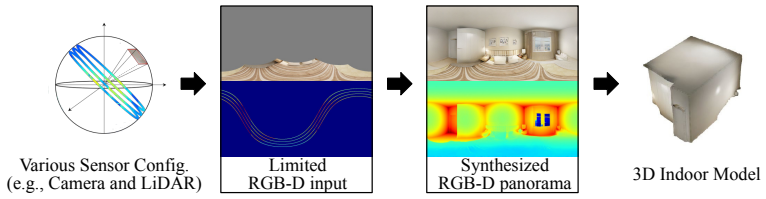

Fig. 1. Overall scheme for our BIPS framework, which takes RGB-D input from cameras and depth sensors in various configurations and synthesizes an RGB-D panorama.

from limited input information can overcome the limitations and generate 3D virtual space with minimal time and effort. Even though prior works have tried to synthesize RGB panorama images from perspective RGB images [20,60], these methods show limited performance on synthesizing panoramas from small partial views and can not be directly extended for RGB-D panorama synthesis.

By contrast, *jointly learning to synthesize depth data along with the RGB images* accompanies two advantages: (1) Depth panorama, which is useful to plenty of applications, can be directly obtained without additional endeavors such as monocular depth estimation or depth completion. (2) The quality of generated RGB and depth panorama can be improved to complement each other. It is because they share the semantic correspondence of the scene, and this correspondence is learned during the joint learning. The extensive experiments in Sec. 4.2 demonstrates the mutual gain between RGB and depth panorama. Therefore, it is promising to synthesize RGB-D panorama from the cameras and depth sensors, such that we can synthesize realistic 3D indoor models.

In this paper, we consider a novel problem: *RGB-D panorama synthesis from limited input visual information of a scene*. To maximize usability, we consider the various configurations of cameras and depth sensors. To this end, we design the various sensor configurations by randomly sampling the number of sensors, their intrinsic parameters, and extrinsic parameters, assuming that the sensors are calibrated and aligned to each other. This enables to represent most of the possible combinations of cameras and depth sensors. Accordingly, our novel bi-modal panorama synthesis (BIPS) framework synthesizes RGB-D indoor panoramas from the camera and depth sensors in various configurations via adversarial learning (See Fig. 3). We thus design a generator that fuses the bi-modal (RGB and depth) features. Through the generator, multiple latent features from one branch can help the other by providing the relevant information of different modalities. For synthesizing the depth of *indoor* scenes, we rely on the fact that the overall layout is usually made of flat surfaces, while interior components have various structures. Thus, we propose to separate the depth of a scene I^{d} into two components: layout depth $I^{\mathrm{d,lay}}$ and residual depth $I^{\mathrm{d,res}}$. Here, $I^{\mathrm{d,lay}}$ corresponds to the depth of planar surfaces, and $I^{\mathrm{d,res}}$ corresponds to the depth of other objects, *e.g.*, furniture. With this relation, we propose a joint learning scheme called *Residual Depth-aided Adversarial Learning (RDAL)*. RDAL

jointly trains RGB panorama, layout depth, and residual depth to synthesize more realistic RGB-D panoramas and 3D indoor models (Sec. 3.2).

Previously, some metrics [56,22] have been proposed to evaluate the outputs of generative models using latent feature distribution of a pre-trained classification network [62]. However, the input modality of utilizing an off-the-shelf network is only limited to perspective RGB images. For this reason, we propose a novel metric, called Fréchet Auto-Encoder Distance (FAED), to evaluate the perceptual quality for RGB-D panorama synthesis (Sec. 3.3). FAED adopts an auto-encoder to reconstruct the inputs from latent features with an unlabeled dataset. Then, the latent feature distribution of the trained auto-encoder is used to calculate the Fréchet distance between the synthesized and GT RGB-D data. Extensive experimental results demonstrate that our RGB-D panorama synthesis method significantly outperforms the extensions of the prior image inpainting [46,88,61], image outpainting [32,60], and image-guided depth synthesis methods [11,51,37,24] modified to synthesize RGB-D panorama from partial RGB-D inputs. Moreover, we show the validity of the proposed FAED by showing how well it captures the disturbance level [22] of synthesized RGB-D panorama.

In summary, our main contributions are three-fold: (I) We introduce a new problem of generating RGB-D panoramas from partial RGB-D inputs under various sensor configurations. (II) We propose a BIPS framework that synthesizes RGB-D panoramas via residual depth-aided adversarial learning. (III) We introduce a novel evaluation metric, FAED, for RGB-D panorama synthesis.

2 Related Works

Image Inpainting. Conventional approaches explore diffusion or patch matching [5,6,7,13,16,8,14]. However, they have limited ability inpainting largely missing regions. The learning-based methods often use generative adversarial networks (GANs) [89,38,26,79], optimized by the minimax loss [27]. Some works explored different convolution layers, *e.g.*, partial convolution [40] and gated convolution [80,50], to handle missing pixels. Moreover, attention mechanism [65,66] has also been applied to capture the contextual information [79,75,41,70,39]. Recently, research has been made to synthesize high-resolution outputs [59,72,52] or semantically diverse outputs [42,87]. Although endeavors have tackled large completion problem [46,88,61], they often fail to synthesize visually pleasing panoramas due to only using perspective RGB inputs.

Image Outpainting. Conventional methods extend an input image to a larger seamless one; however, they require manual guidance [4,6,86] or image sets of the same scene category [29,58,69]. By contrast, learning-based methods synthesize large images with novel textures that do not exist in the input image [55,34,74,18,30,82,47,31,19]. Some works focus on driving scenes [73,84] or synthesize panorama with iterative extension or multiple perspective images [78,32,20,60]. Although performance has been greatly improved, the existing methods are still afflicted by the limited quality from the perspective images.

Image-guided Depth Synthesis. One line of research attempts to fuse the bi-modal information, *i.e.*, the RGB image and sparse depth. Some methods, *e.g.* [44], fuse the sparse depth and RGB image via early fusion while others [43,28,63,17,37,25] utilize a late fusion scheme, or jointly utilize both [64,36,67]. Another line of research focuses on utilizing affinity or geometric information of the scene via surface normal, occlusion boundaries, and the geometric convolutional layer [33,53,76,85,24,11,10,51]. However, these works only generate dense depth maps that have the same FoV with the input perspective RGB images.

Evaluation of Generative Models. Image quality assessment can be classified into three groups: full-reference (FR), reduced-reference (RR), and no-reference (NR). There exist many conventional FR metrics, *e.g.*, PSNR, MSE, and SSIM, and deep learning (DL)-based FR metrics, *e.g.*, LPIPS [83]. These metrics typically calculate either pixel-wise or patch-wise similarity to the ground truth images. By contrast, NR methods, *e.g.*, BRISQUE [48] and NIQE [49] do not require reference image. Among the DL-based NR metrics, Inception Score (IS) [56] and Fréchet Inception Distance (FID) [22] are widely used [2]. IS and FID scores are calculated from pretrained classification models to capture the high-level features. Unfortunately, these metrics are less applicable for RGB-D panorama evaluation because (1) they are trained only with perspective RGB images, and (2) there are no labeled panorama images for training. They are highly sensitive to the distortion of panoramas, making them hard to capture perceptual quality properly on panoramas. Furthermore, naively using them on RGB-D leads to an imprecise measure of the semantic correspondence between the two different modalities. Therefore, *we propose FAED, which aims to directly evaluate the RGB-D panorama quality. FAED can be adaptively applied to evaluate multi-modal data that lacks a labeled dataset.*

3 Proposed Methods

3.1 Problem Formulation

Previous works, *e.g.*, [60,20], generate an equirectangular projection (ERP) image (ERP^{rgb}) from input RGB image(s) ($I_{\mathrm{in}}^{\mathrm{rgb}}$). Then, an RGB panorama $I_{\mathrm{out}}^{\mathrm{rgb}}$ can be created via a function G, mapping $I_{\mathrm{in}}^{\mathrm{rgb}}$ into a ERP^{rgb} [21], which can be formulated as $I_{\mathrm{out}}^{\mathrm{rgb}} = ERP^{\mathrm{rgb}} = G(I_{\mathrm{in}}^{\mathrm{rgb}})$.

However, as it is crucial to provide omnidirectional depth information [54,1] in many applications, many studies tried to synthesize depth panoramas from input RGB panorama and partial depth measurements [68,23]. One solution to synthesize an RGB-D panorama would be to sequentially synthesize RGB panorama from input RGB images, and then apply the depth synthesis methods to generate an omnidirectional depth map. However, such an approach is cumbersome and less effective, as shown in the experimental results (See Table 4). We solve this novel yet challenging problem by jointly utilizing the input RGB image ($I_{\mathrm{in}}^{\mathrm{rgb}}$) and depth data ($I_{\mathrm{in}}^{\mathrm{d}}$). Our goal is to directly generate the RGB panorama (ERP^{rgb}) and depth panorama (ERP^{d}) simultaneously via a mapping function

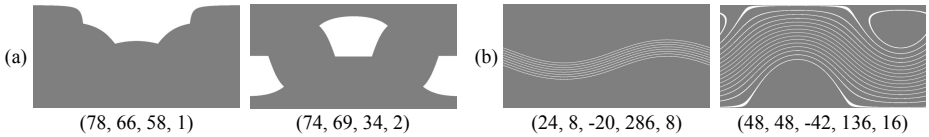

Fig. 2. Sampled input masks. (a) Input mask of cameras and perspective depth sensors with parameters $(\delta_H, \delta_V, \psi, n)$ and (b) mechanical LiDARs with $(\delta_L, \delta_U, \psi, \omega, \eta)$.

G, which can be described as $(I^{\text{rgb}}_{\text{out}}, I^{\text{d}}_{\text{out}}) = (ERP^{\text{rgb}}, ERP^{\text{d}}) = G(I^{\text{rgb}}_{\text{in}}, I^{\text{d}}_{\text{in}})$. G can be formulated by learning a *single* network to synthesize ERP^{rgb} and ERP^{d} using $I^{\text{rgb}}_{\text{in}}$ and I^{d}_{in} obtained in various sensor configurations. As the information in the left and right boundaries in ERP images should be connected, our designed G uses circular padding [57] before each convolutional operation.

Consequently, we design the various input configurations by randomly sampling the parameters of cameras and depth sensors to provide the input to the G during training. Since our model takes partial ERP input, data obtained from sensors should be projected to ERP image space. The masks of the sensor input area projected on the ERP can be represented as shown in Fig. 2. Therefore, the partial RGB-D input projected on the ERP space, $I^{\text{rgb}}_{\text{in}}$ and I^{d}_{in}, can be obtained by multiplying sampled mask to the full ERP image. We also randomly choose whether to use cameras only, depth sensors only, or both, to handle the cases where only cameras or depth sensors are available.

Parameters of RGB Cameras. We denote the parameters of RGB cameras, horizontal FoV as δ_H, vertical FoV as δ_V, pitch angle as ψ, and the number of viewpoints as n. When $n > 1$, we arrange the viewpoints in a circle having the sampled pitch angle from the equator and at the same intervals. We do not consider roll and yaw, as they do not affect the results (*i.e.*, the output is equivariant to the horizontal shift of input) thanks to using circular padding. Considering general setting of cameras, we sample the parameters from $\{\delta_H, \delta_V\} \sim \mathcal{U}[60°, 90°]$, $\psi \sim \mathcal{U}[-90°, 90°]$, and $n \sim \mathcal{U}\{0, 1, 2, 3, 4\}$, where $\mathcal{U}[\cdot]$ represents uniform distribution.

Parameters of Depth Sensors. I^{d}_{in} can be obtained from mechanical LiDARs or perspective depth sensors thus we should generate various depth input masks for both. For the LiDARs, we denote lower FoV as δ_L, upper FoV as δ_U, pitch angle as ψ, yaw angle as ω, and the number of channels as η. The yaw angle is needed to consider the relative yaw motion to the camera arrangement. For the perspective depth sensors providing dense depth, we use the same sampled parameters with the cameras $(\delta_H, \delta_V, \psi, n)$. In practice, we first sample the parameters from $\psi \sim \mathcal{U}[-90°, 90°]$, $\omega \sim \mathcal{U}[0°, 360°]$, and $\eta \sim \mathcal{U}\{0, 2, 4, 8, 16\}$. Then, we sample δ_L and δ_U from $\mathcal{U}\{\eta, 2\eta, 3\eta\}$. Finally, our problem is formulated as

$$(I^{\text{rgb}}_{\text{out}}, I^{\text{d}}_{\text{out}}) = (ERP^{\text{rgb}}, ERP^{\text{d}}) \tag{1}$$

$$= G(I^{\text{rgb}}_{\text{in}}(\delta_H, \delta_V, \psi, n), I^{d}_{in}(\delta_L, \delta_U, \psi, \omega, \eta, \delta_H, \delta_V, n)) \tag{2}$$

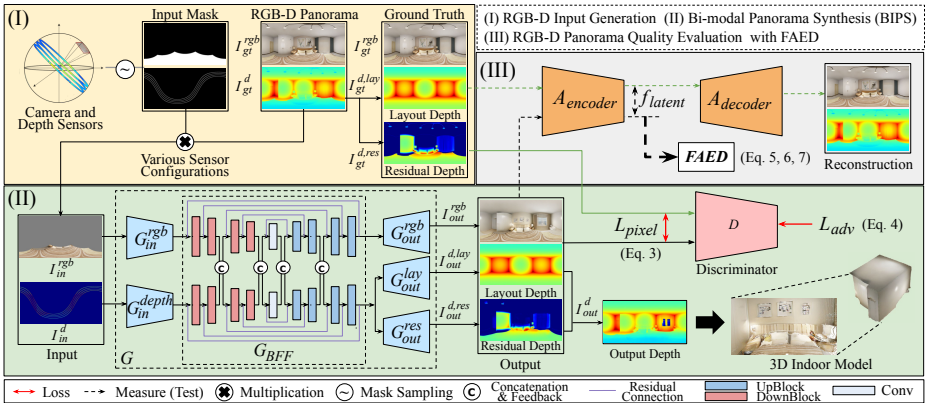

Fig. 3. Overall structure of our bi-modal indoor panorama synthesis (BIPS) framework. Our framework takes RGB-D input provided by various sensor configurations, integrates the bi-modal input data with BFF branch in the generator network, and jointly trains to synthesize layout depth and residual depth. Then, the perceptual quality of the synthesized RGB-D panorama is measured by proposed FAED metric.

3.2 RGB-D Panorama Synthesis Framework

Overview. An overview of the proposed BIPS framework is depicted in Fig. 3. BIPS consists of a generator G, and a discriminator D. G takes the partial RGB image $I_{\text{in}}^{\text{rgb}}$ and depth I_{in}^{d} as inputs. We notice that the quality of the RGB-D panorama depends on both the overall (mostly rectangular) layout and how the furniture are arranged in the indoor scene. Inspired by [81], we separate the depth data I_{gt}^{d} into *layout depth* $I_{\text{gt}}^{\text{d,lay}}$, and *residual depth (interior components)* $I_{\text{gt}}^{\text{d,res}}$, which is defined as ($I_{\text{gt}}^{\text{d}}$ - $I_{\text{gt}}^{\text{d,lay}}$). The generator G outputs the RGB panorama $I_{\text{out}}^{\text{rgb}}$, the layout depth panorama $I_{\text{out}}^{\text{d,lay}}$, and the residual depth panorama $I_{\text{out}}^{\text{d,res}}$ simultaneously. As these are jointly trained with adversarial loss, we call this learning scheme *Residual Depth-aided Adversarial Learning (RDAL)*.

Generator. G is composed of input branch G_{in}, bi-modal feature fusion (BFF) branch G_{BFF}, and output branch G_{out}, as shown in Fig. 3. G_{in} consists of two encoding branches: $G_{\text{in}}^{\text{rgb}}$ and $G_{\text{in}}^{\text{depth}}$, which take $I_{\text{in}}^{\text{rgb}}$ and I_{in}^{d} respectively, that independently extract RGB and depth features. Then, G_{BFF} fuses the highly correlated features to fully exploit the bi-modal information of the scene. Lastly, G_{out} have three decoding branches and each of them generates RGB panorama $I_{\text{out}}^{\text{rgb}}$, layout depth panorama $I_{\text{out}}^{\text{d,lay}}$, and residual depth panorama $I_{\text{out}}^{\text{d,res}}$, respectively. Since realistic indoor space comes with clean layout structure and detailed interiors, we design G_{out} to jointly synthesize the layout and residual depth of an indoor scene. The detailed structure can be found in suppl. material.

BFF Branch. G_{BFF} takes $G_{\text{in}}^{\text{rgb}}(I_{\text{in}}^{\text{rgb}})$ and $G_{\text{in}}^{\text{depth}}(I_{\text{in}}^{\text{d}})$ as inputs. Since $I_{\text{in}}^{\text{rgb}}$ and I_{in}^{d} are captured in the same scene, their bi-modal information is highly correlated in a semantic manner. To utilize this correlation, G_{BFF} consists of two-stream

encoder-decoder networks fusing the bi-modal features. The bi-modal features are fused in between the layers of G_{BFF} four times. In particular, the features from both branches are concatenated and fed back to each other. Overall, the fusion is done after the features pass two 'DownBlocks' and before the features pass two 'UpBlocks'. The 'UpBlock' consists of one convolution layer with 4×4 kernel and three convolution layers with 3×3 kernel. The 'DownBlock' consists of one upsample layer, one convolution layer with 4×4 kernel and three convolution layers with 3×3 kernel. In addition, multi-scale residual connections are used to vitalize the transfer of information between the layers and branches. As multiple latent features from one branch help the other by sharing the information apart in both ways, G_{BFF} can generate features by fully exploiting the information of the scene.

Discriminator. We use the multi-scale discriminator D from [71], but modify it to have five input channels (three for I^{rgb}, one for $I^{\mathrm{d,lay}}$, and one for $I^{\mathrm{d,res}}$). The detailed discriminator structure can be found in the suppl. material.

Loss Function. For training G, we use a weighted sum of the pixel-wise L1 loss and adversarial loss. The pixel-wise L1 loss between the GT and the output panorama, denoted as L_{pixel}, consists of three terms as the G has three outputs (RGB, layout depth, residual depth panorama):

$$L_{\mathrm{pixel}}^{\mathrm{total}} = L_{\mathrm{pixel}}^{\mathrm{rgb}} + L_{\mathrm{pixel}}^{\mathrm{d,lay}} + L_{\mathrm{pixel}}^{\mathrm{d,res}}. \tag{3}$$

For adversarial loss L_{adv}, we used LSGAN loss [45]: $L_{\mathrm{adv}} = \frac{1}{2}\,\mathbb{E}\,[(D(I_{\mathrm{out}}^{\mathrm{total}}) - 1)^2]$, where $I_{\mathrm{out}}^{\mathrm{total}}$ is the concatenation of generator outputs $I_{\mathrm{out}}^{\mathrm{rgb}}$, $I_{\mathrm{out}}^{\mathrm{d,lay}}$ and $I_{\mathrm{out}}^{\mathrm{d,res}}$, and D is a discriminator. By decomposing the total depth loss into $L^{\mathrm{d,lay}}$ and $L^{\mathrm{d,res}}$, our RDAL scheme allows G to synthesize RGB-D panorama that has a highly plausible interior. Finally, the total loss for the generator is:

$$L_{\mathrm{G}} = \lambda L_{\mathrm{pixel}}^{\mathrm{total}} + L_{\mathrm{adv}} \tag{4}$$

where λ is a weighting factor. For detailed loss terms, refer to suppl. material.

3.3 Fréchet Auto-Encoder Distance (FAED)

Auto-Encoder Network. Similar to the high-level features in a CNN trained with large-scale semantic labels, latent features f_{latent} in a trained auto-encoder also contain high-level information, as auto-encoder is trained to reconstruct the input from the latent features. Moreover, the auto-encoder has the advantage that it does not need any labels for training. Since there is no dataset including semantic labels for RGB-D panoramas, we propose to train an auto-encoder A to generate RGB-D panoramas and use its latent features to calculate the perceptual quality. The detailed structure of A is given in the suppl. material.

Calculation of FAED for RGB-D Panorama. We denote f_{latent} at c-th channel, h-th row, and w-th column as $f_{\mathrm{latent}}(c, h, w)$. Note that as we use ERP, the h and w have one-to-one relation to latitude and longitude.

Longitudinal Invariance. To evaluate the performance of G, we extract f_{latent} from generated samples using A_{encoder}. However, as we generate the upright ERP

Fig. 4. The proposed FAED metric for RGB-D panorama quality evaluation. It measures the distance of the distributions of latent features extracted from the pre-trained auto-encoder network on RGB-D panorama.

image, it is expected to have a distance metric that is invariant to the longitudinal shift. This is because an upright ERP panorama represents the same scene when it is cyclically shifted along the longitudinal direction. Therefore, to make the resulting distance metric invariant to the longitudinal shift, we take the mean along the longitudinal direction of f_{latent} as:

$$f'_{\text{latent}}(c,h) = \frac{1}{W}\sum_{w} f_{\text{latent}}(c,h,w). \tag{5}$$

Latitudinal Equivariance. As ERP has varying sampling rates depending on the latitude ϕ, we apply different weights on f'_{latent} considering information density along latitude. Specifically, we multiply $\cos(\phi)$ to feature at the latitude ϕ, since in ERP, each pixel occupies $\cos(\phi)$ area in the spherical surface, compared with the pixels in the equator. The resulting feature is expressed as:

$$f''_{\text{latent}}(c,h) = \cos(\phi)\cdot f'_{\text{latent}}(c,h). \tag{6}$$

Fréchet Distance. We treat the resulting f''_{latent} as a vector and assume that it has a multi-dimensional Gaussian distribution. Then, we get the distribution of ground truths $\mathcal{N}(m,C)$ and that of generated samples $\mathcal{N}(\hat{m},\hat{C})$, and calculate the Fréchet distance d between them as given by [15]:

$$d^2(\mathcal{N}(m,C),\mathcal{N}(\hat{m},\hat{C})) = ||m-\hat{m}||_2^2 + Tr(C+\hat{C}-2(C\hat{C})^{1/2}). \tag{7}$$

We use d^2 as a perceptual distance metric where m and C is mean and covariance.

4 Experimental Results

Synthetic Dataset. Structured3D dataset [90] provides various textures of indoor scenes with a 512×1024 resolution. We split the dataset into 17468 train, 2183 validation, and 2184 test data. Then we augmented the entire data with three random horizontal shifts then the number of the dataset has quadrupled. In addition, with the corner locations provided in the dataset, we manually generated layout depth maps of each 3D scene. The residual depth maps are

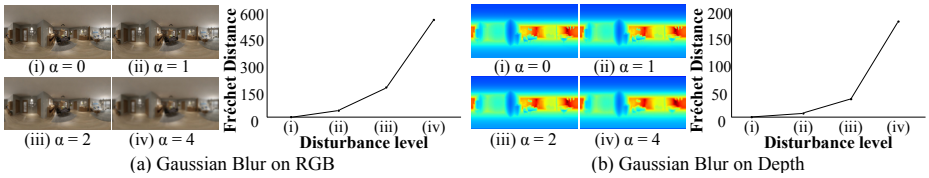

Fig. 5. Verification of FAED in Structured3D dataset. It can be seen that FAED correlates well with perceptual evaluation of humans, as FAED increases as the data becomes more corrupted. For more detailed results, please refer to the suppl. material.

obtained by subtracting the layout depth from the GT depth map. More details about GT layout and residual depth generation can be found in suppl. material.
Real Dataset. We used a combination of two datasets: Matterport3D [9] and 2D-3D-S dataset [3]. Both datasets provide real-world indoor RGB-D panorama captured with real sensors, so depth data in those datasets contain sensor noise or missing holes. However, since this dataset does not provide a sufficient number of annotated layouts, it is only used for test purpose.
Implementation Details. Please refer the suppl. material. for the details.

4.1 Verification of FAED

To show the effectiveness of FAED on measuring the perceptual quality of RGB-D panorama, we corrupt the Structured3D dataset [90] in two ways: corrupting RGB only or corrupting depth only. Following [22], we corrupt the dataset by applying various types of noise: Gaussian blur, Gaussian noise, uniform patches, swirl, and salt and pepper noise. Also, we utilized discrete cosine transform that causes blocking effects to show that our model is sensitive to GAN-like artifacts. Here, we plot the result of Gaussian blur in Fig. 5. Other results can be found in suppl. material. *Note that the evaluation is done for RGB-D panorama, neither for RGB image alone nor for depth map alone.* As shown in Fig. 5, the Fréchet distance for both RGB and depth panorama increases as the disturbance level is increased. We show that the same applies to the other five types of noises in the suppl. material due to the lack of space. *This demonstrates the perceptual quality of RGB-D panorama becomes poorer as the FAED increases.*

Moreover, we calculated FAED of paired and unpaired RGB-D panorama to verify that FAED is effective in considering semantic alignment between RGB and depth panorama. Unpaired RGB-D panorama consists of RGB panorma and randomly selected depth panorama not corresponded to RGB panorama, and its FAED score is 168.0. 3D indoor model from unpaired RGB-D panorama has inconsistent semantic information, *e.g.*, misaligned corner of indoor room and distorted furniture. The visual results of the inconsistency can be found in suppl. material. *Consequently, it indicates that the higher FAED score denotes poorer semantic alignment between RGB and depth panorama.*

Table 1. Quantitative results of RGB panorama synthesis on Structured3D dataset. As [60] needs 4 perspective RGB inputs, we report our results in the same setting. In other cases, 1∼4 number of RGB inputs are randomly used. The depth input is not used to compare with image synthesis methods. For FAED calculation, GT full depth is used with synthesized RGB panorama. Best results in **bold**.

Category	Method	Input no. (n)		RGB metric			Layout metric	Proposed metric
		RGB	Depth	PSNR(↑)	SSIM(↑)	LPIPS(↓)	2D Corner error(↓)	FAED(↓)
Inpainting	BRGM [46]	1/2/3/4	0	14.00	0.5310	0.6192	72.52	442.3
	CoModGAN [88]			14.35	0.5837	0.4768	62.45	208.2
	LaMa [61]			13.74	0.5207	0.5658	51.12	379.2
Outpainting	Boundless [32]			13.74	0.5663	0.6144	74.47	429.4
	Ours			**16.21**	**0.6161**	**0.4549**	**39.63**	**162.3**
Panorama syn.	Sumantri *et. al.*[60]	4	0	**18.49**	**0.6680**	0.4190	50.76	443.4
	Ours			17.29	0.6510	**0.3975**	**34.68**	**103.1**

Fig. 6. Qualitative comparison to Sumantri *et. al.* [60]. While the result from [60] is blurry, our result is sharp and realistic.

4.2 RGB-D Panorama Synthesis

Evaluation on RGB Panorama Synthesis. Table 1 shows the quantitative comparison with the inpainting and outpainting methods on the Structured3D dataset. Our model takes partial RGB inputs and no depth as inputs. We use PSNR, SSIM, and LPIPS to evaluate the quality of RGB panorama. We also measure 2D corner error, where the 2D GT corner points are compared with the estimated 2D corner points using DuLa-Net [77] on the synthesized RGB panorama. We also use the proposed FAED to evaluate the perceptual quality. Here, synthesized RGB panorama and GT depth are used to compute FAED.

As shown in Table 1, our method outperforms the image inpainting and outpainting methods: BRGM [46], CoModGAN [88], LaMa [61], and Boundless [32], by a large margin for all metrics. For instance, our method outperforms the best inpainting method, CoModGAN, by a 4.6% decrease in LPIPS score, 36.5% drop of 2D corner error, and 22% decline of FAED score. The effectiveness can also be visually verified in Fig. 7(a). Our method produces clearer RGB panorama images compared with LaMa producing blurry images. Although CoModGAN produces clear RGB outputs, it does not consider the indoor layout and semantic information of the furniture, *e.g.*, the electric cooker is combined with bookshelves, as shown in Fig. 7. The reason our model has higher performance than the existing SoTA inpainting/outpainting method is that RDAL helps the generator to learn distinguishing features of layout and residual during joint learning. Although the layout and the residual are separated only in the depth image, our joint learning framework induces learning of highly correlated features

Table 2. Quantitative results of depth panorama synthesis on Structured3D dataset. Depth input type L/P means that we use LiDAR (L) and dense perspective depth sensor (P). The full GT RGB is used with synthesized depth panorama for FAED calculation. Best results in **bold**.

Category	Method	Input type		Depth metric		Layout metric	Proposed metric
		RGB	Depth	AbsREL(↓)	RMSE(↓)	2D IoU(↑)	FAED(↓)
Depth syn.	CSPN [11]	Full	L/P	0.0855	2214	0.8062	428.9
	NLSPN [51]			0.1268	2807	0.7333	836.1
	MSG-CHN [37]			0.1764	3296	0.6724	896.4
	PENet [24]			0.1740	3145	0.7033	906.0
	Ours			**0.0844**	**1942**	**0.8286**	**131.5**

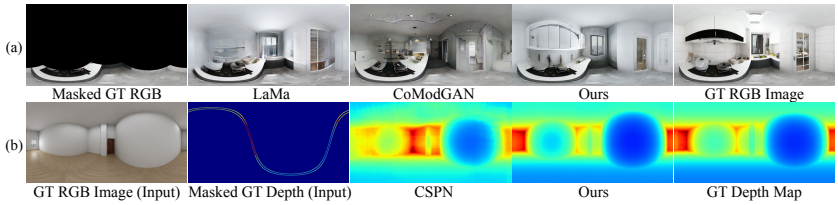

Fig. 7. (a) Visual results for RGB panorama synthesis on Structured3D dataset. Two methods, LaMa and CoMoGAN, are visualized for comparison. (b) Visual results for depth panorama synthesis on Structured3D dataset. CSPN is also visualized for comparison. More qualitative results can be found in suppl. material.

by exchanging information between depth and RGB. Therefore, it is possible to create a very realistic indoor environment panorama even in RGB compared to other models that do not take this into account.

We also compare our model with the panorama synthesis method, Sumantri *et. al.* [60]. Our method shows slightly lower scores using the conventional metrics, PSNR and SSIM; however, it shows a much better LPIPS score, 2D corner error, and FAED score as shown Table 1. We argue that PSNR and SSIM merely measure local photometric similarity and thus fail to reflect the perceptual quality while FAED catches the perceptual quality. This can be visually verified in Fig. 6. Our method synthesizes better textures and shows much more visually plausible output. More visual results can be found in suppl. material.

Evaluation on Depth Panorama Synthesis. We compare our method with the image-guided depth synthesis methods: CSPN [11], NLSPN [51], MSG-CHN [37] and PENet [24]. AbsREL, RMSE, layout 2D IoU [12] and the proposed FAED are used for evaluation. The details of the metrics can be found in the suppl. material. Table 2 shows the quantitative comparison with the depth synthesis methods. In particular, our method outperforms one of the best depth synthesis method, CSPN, with all metrics. We also compared our model with the 360° monocular depth estimation method [67]. The visual result of [67] was not reasonable, and its FAED score was 1140. With the proposed RDAL scheme, our model understands the structure of the indoor scene and learns the relative

Table 3. FAED scores of our model according to the amount of RGB-D inputs.

FAED (The redder the cell, the lower the value)		Number of Depth Input								
			Perspective (num. of NFoVs)				LiDAR (num. of channels)			
		0	1	2	3	4	2	4	8	16
Number of Perspective RGB Input	0	-	2077	746.4	439.8	371.0	1345	1030	631.1	382.6
	1	910.0	695.1	246.6	176.5	152.2	316.6	267.9	210.6	151.7
	2	461.8	295.4	202.3	152.2	132.7	229.4	207.8	174.6	134.4
	3	365.8	233.7	154.0	128.3	107.5	189.9	171.4	141.7	101.4
	4	346.1	214.5	141.7	108.0	91.9	176.4	156.3	127.9	87.2

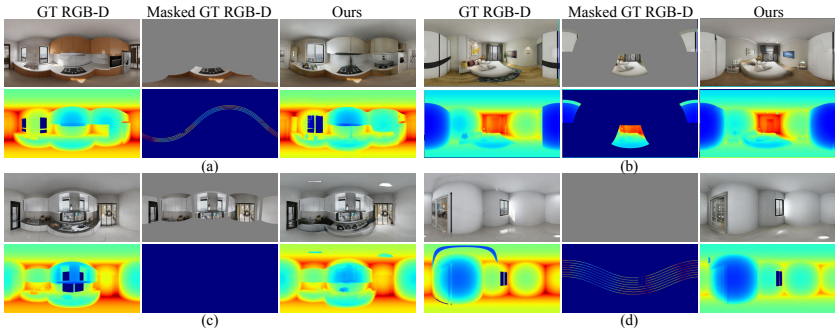

Fig. 8. Visualization of our synthesized RGB-D panorama results from RGB-D data in various configurations. (a) and (b) take both RGB and depth data, (c) takes only RGB, and (d) takes only depth data. More results are visualized in suppl. materials.

depth of interior components. Therefore, our method estimates the best layout depth, which is demonstrated by the highest layout 2D IoU.

Fig. 7(b) shows the qualitative comparison with CSPN [11]. CSPN failed to synthesize valid layouts with non-planar output depth map on the walls and ceiling, which incurs unrealistic 3D indoor model. In contrast, our result shows undisturbed, clear layouts. More of these results can be found in suppl. material.

Evaluation on RGB-D Panorama Synthesis. To show the effectiveness of our model quantitatively, we compared our model with 'inpainting with depth synthesis' (IwDS). To be specific, an RGB panorama is first synthesized from partial RGB input using the image inpainting method. Then, depth panorama is synthesized by applying the depth synthesis method to the synthesized RGB panorama and partial depth input. We chose CoModGAN [88] and CSPN [11] for RGB and depth synthesis methods, which showed the highest FAED score in Table 1 and Table 2. In Table 4, it can be seen that IwDS leads lower 2D IoU score and a much higher FAED score than our method. Also, FAED score of IwDS with [60] and [11] was 722.1, even though [60] uses 1∼4 RGB inputs. These indicate that the two-stage, sequential synthesis of RGB-D panorama is less effective than our BIPS framework that fuses the bi-modal features, trained with one-stage, joint learning scheme. Also, IwDS fails to generate realistic 3D indoor models, with distorted indoor layouts and severe bumpy surfaces, as shown in Fig. 10.

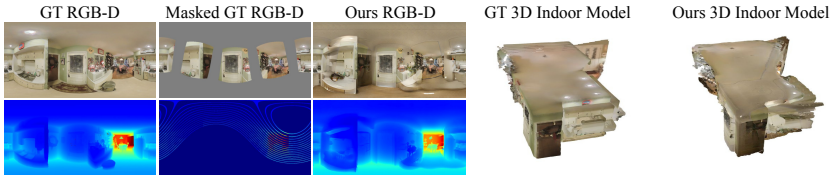

Fig. 9. Visual results for RGB panorama synthesis on Matterport3D dataset.

Fig. 8 shows the qualitative results of our model generated from the partial RGB and depth inputs, including RGB only or depth only cases. The mutual gain for using RGB and depth information together is visually demonstrated in Fig. 8(a). The upper shelf not visible in RGB input is successfully generated by utilizing the corresponding depth data. Likewise, the lower shelf not visible from the depth input is plausibly created in output depth panorama referring to RGB information. It means that the bi-modal information is exchanged in a bi-directional manner, which enables our model to understand the overall scene. *More results with 3D indoor models are visualized in suppl. materials.*

In Table 3, we quantitatively analyze the performance with FAED regarding the amount of input information. Overall, *using both types of input shows better panorama synthesis quality than that of using a single input.* For example, using 2 RGB and 2 depth inputs shows much better FAED score (202.3) than those of using 4 RGB or 4 depth inputs (FAED scores: 371.0, 346.1). The fusion between textural information from RGB and structural information from depth through BFF enables a more comprehensive understanding of the scene.

RGB-D Panorama Evaluation on Real Dataset. We evaluated our synthesized RGB-D panorama on real indoor scenes in Matterport3D and 2D-3D-S dataset. Fig. 9 shows an output RGB-D panorama and its 3D indoor model. Overall, our method synthesizes high-quality RGB-D panorama on real indoor scenes, which are unseen during training. Our synthesized depth panorama shows the precise indoor layout and plausible residuals, generating a realistic 3D indoor model. For quantitative results, our method achieved a much better FAED score than IwDS (1123 vs. 5099). Since the domain of real dataset is different from the domain of synthetic dataset, FAED scores generally increased. More visual results can be found in suppl. material.

4.3 Ablation Study and Analysis

Impact of BFF. We studied the effectiveness of RGB-D panorama synthesis by removing the BFF branch in the generator. In details, G_{BFF} is replaced with a single branch network taking the concatenation of $G_{\text{in}}^{\text{rgb}}(I_{\text{in}}^{\text{rgb}})$ and $G_{\text{in}}^{\text{depth}}(I_{\text{in}}^{\text{d}})$. For fair comparison, we designed a single branch network to be of similar capacity to our final model (23,254 vs. 22,642 MB). As shown in Table 4, the 2D IoU drops and FAED increases without BFF. Fig. 10 shows that the texture of the RGB-D output is not consistent with the given RGB-D input. This reflects that BFF

Table 4. Quantitative results of IwDS and ablation study of BIPS framework.

	IwDS ([88]+[11])	Ours w/o BFF	Ours w/o RDAL	Ours
2D IoU(↑)	0.7561	0.7859	0.7164	**0.8158**
FAED(↓)	640.9	381.4	329.0	**198.0**

Fig. 10. Visualization of IwDS and our ablation study results on Structured3D dataset.

encourages the information exchange of bi-modal information and significantly contributes to having minimal artifacts.

Impact of RDAL. We further compared the model without RDAL to validate its effectiveness. In detail, the number of output branches is reduced to two, and each is designed to learn RGB and total depth panorama. As shown in Table 4, the 2D IoU drops and FAED increases without RDAL. It shows that RDAL is critical for estimating precise indoor layouts. The impact of RDAL is visually verified in Fig. 10. The result without RDAL shows a distorted indoor layout while having fewer artifacts than ours without BFF. In summary, jointly learning layout and residual depth helps to synthesize a more structural 3D indoor model.

Analysis on Robustness of BIPS. We conducted various experiments: applying on noisy sensor inputs, comparison on different input data of the same 3D scene, and generalization on unseen input configurations. The results show that the proposed BIPS can synthesize visually pleasing RGB-D panorama under these scenarios, making it directly applicable to real-world applications. The implementation details, results, and discussion are included in suppl. material.

5 Conclusion

We tackled a novel problem of synthesizing RGB-D indoor panoramas from various configurations of RGB and depth inputs. Our method can synthesize high-quality RGB-D panoramas with the proposed BIPS framework by utilizing the BFF and jointly training through RDAL. Extensive experiments show that this bi-modal joint learning enables the generator to effectively understand the structure of indoor scene, so our model achieved the highest performance in indoor RGB-D panorama synthesis than conventional methods. Moreover, a label-free novel image quality assessment metric, FAED, was proposed, and its validity was demonstrated. **Acknowledgements.** This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636). This research was conducted while Wonjune Cho and Lin Wang were with KAIST.

References

1. Alaee, G., Deasi, A.P., Pena-Castillo, L., Brown, E., Meruvia-Pastor, O.: A user study on augmented virtuality using depth sensing cameras for near-range awareness in immersive vr. In: IEEE VR's 4th Workshop on Everyday Virtual Reality (WEVR 2018). vol. 10 (2018)
2. Ali, B.: Pros and cons of gan evaluation measures. Computer Vision and Image Understanding **179**, 41–65 (2019)
3. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
4. Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. In: ACM SIGGRAPH 2007 Papers. p. 10–es. SIGGRAPH '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1275808.1276390
5. 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)
6. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
7. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 417–424 (2000)
8. Bertalmio, M., Vese, L., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. IEEE transactions on image processing **12**(8), 882–889 (2003)
9. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017)
10. Cheng, X., Wang, P., Chenye, G., Yang, R.: Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. Proceedings of the AAAI Conference on Artificial Intelligence **34**, 10615–10622 (04 2020). https://doi.org/10.1609/aaai.v34i07.6635
11. Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2361–2379 (2020). https://doi.org/10.1109/TPAMI.2019.2947374
12. Choi, D.: 3d room layout estimation beyond the manhattan world assumption. arXiv preprint arXiv:2009.02857 (2020)
13. Criminisi, A., Perez, P., Toyama, K.: Object removal by exemplar-based inpainting. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 2, pp. II–II. IEEE (2003)
14. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on image processing **13**(9), 1200–1212 (2004)
15. Dowson, D., Landau, B.: The fréchet distance between multivariate normal distributions. Journal of multivariate analysis **12**(3), 450–455 (1982)
16. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2, pp. 1033–1038. IEEE (1999)
17. Eldesokey, A., Felsberg, M., Khan, F.S.: Confidence propagation through cnns for guided sparse depth regression. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2423–2436 (Oct 2020).

https://doi.org/10.1109/tpami.2019.2929170, http://dx.doi.org/10.1109/TPAMI.2019.2929170
18. Guo, D., Feng, J., Zhou, B.: Structure-aware image expansion with global attention. In: SIGGRAPH Asia 2019 Technical Briefs. p. 13–16. SA '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3355088.3365161
19. Guo, D., Liu, H., Zhao, H., Cheng, Y., Song, Q., Gu, Z., Zheng, H., Zheng, B.: Spiral generative network for image extrapolation. In: Computer Vision — ECCV 2020. pp. 701–717. Springer-Verlag, Berlin, Heidelberg (2020)
20. Hara, T., Harada, T.: Spherical image generation from a single normal field of view image by considering scene symmetry. Proceedings of the AAAI Conference on Artificial Intelligence **35(2)**, 1513–1521 (05 2021)
21. He, Y., Ye, Y., Hanhart, P., Xiu, X.: Geometry padding for motion compensated prediction in 360 video coding. In: 2017 Data Compression Conference (DCC). pp. 443–443. IEEE Computer Society (2017)
22. 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. In: Advances in neural information processing systems. pp. 6626–6637 (2017)
23. Hirose, N., Tahara, K.: Depth360: Monocular depth estimation using learnable axisymmetric camera model for spherical camera image. CoRR **abs/2110.10415** (2021), https://arxiv.org/abs/2110.10415
24. Hu, M., Wang, S., Li, B., Ning, S., Fan, L., Gong, X.: Penet: Towards precise and efficient image guided depth completion. In: IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021. pp. 13656–13662. IEEE (2021). https://doi.org/10.1109/ICRA48506.2021.9561035
25. Huang, Z., Fan, J., Cheng, S., Yi, S., Wang, X., Li, H.: Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion. IEEE Transactions on Image Processing **29**, 3429–3441 (2020). https://doi.org/10.1109/TIP.2019.2960589
26. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (ToG) **36**(4), 1–14 (2017)
27. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
28. Jaritz, M., De Charette, R., Wirbel, E., Perrotton, X., Nashashibi, F.: Sparse and dense data with cnns: Depth completion and semantic segmentation. In: 2018 International Conference on 3D Vision (3DV). pp. 52–60. IEEE (2018)
29. Kaneva, B., Sivic, J., Torralba, A., Avidan, S., Freeman, W.T.: Infinite images: Creating and exploring a large photorealistic virtual space. Proceedings of the IEEE **98**(8), 1391–1407 (2010). https://doi.org/10.1109/JPROC.2009.2031133
30. Kasaraneni, S.H., Mishra, A.: Image completion and extrapolation with contextual cycle consistency. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 1901–1905 (2020). https://doi.org/10.1109/ICIP40778.2020.9191339
31. Kim, K., Yun, Y., Kang, K.W., Kong, K., Lee, S., Kang, S.J.: Painting outside as inside: Edge guided image outpainting via bidirectional rearrangement with progressive step learning. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 2121–2129 (2021)
32. Krishnan, D., Teterwak, P., Sarna, A., Maschinot, A., Liu, C., Belanger, D., Freeman, W.: Boundless: Generative adversarial networks for image extension. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10520–10529 (2019). https://doi.org/10.1109/ICCV.2019.01062

33. Lee, B., Jeon, H., Im, S., Kweon, I.S.: Depth completion with deep geometry and context guidance. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 3281–3287 (2019). https://doi.org/10.1109/ICRA.2019.8794161
34. Lee, D., Yun, S., Choi, S., Yoo, H., Yang, M.H., Oh, S.: Unsupervised holistic image generation from key local patches. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 21–37. Springer International Publishing, Cham (2018)
35. Lee, J.K., Yea, J., Park, M.G., Yoon, K.J.: Joint layout estimation and global multi-view registration for indoor reconstruction. In: Proceedings of the IEEE international conference on computer vision. pp. 162–171 (2017)
36. Lee, S., Lee, J., Kim, D., Kim, J.: Deep architecture with cross guidance between single image and sparse lidar data for depth completion. IEEE Access **8**, 79801–79810 (2020). https://doi.org/10.1109/ACCESS.2020.2990212
37. Li, A., Yuan, Z., Ling, Y., Chi, W., Zhang, S., Zhang, C.: A multi-scale guided cascade hourglass network for depth completion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (March 2020)
38. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3911–3919 (2017)
39. Liao, L., Xiao, J., Wang, Z., Lin, C.W., Satoh, S.: Image inpainting guided by coherence priors of semantics and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6539–6548 (2021)
40. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 85–100 (2018)
41. Liu, H., Jiang, B., Xiao, Y., Yang, C.: Coherent semantic attention for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4170–4179 (2019)
42. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J.: Pd-gan: Probabilistic diverse gan for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9371–9381 (2021)
43. Ma, F., Cavalheiro, G.V., Karaman, S.: Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 3288–3295 (2019). https://doi.org/10.1109/ICRA.2019.8793637
44. Mal, F., Karaman, S.: Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1–8. IEEE (2018)
45. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017)
46. Marinescu, R.V., Moyer, D., Golland, P.: Bayesian image reconstruction using deep generative models. In: Advances in Neural Information Processing Systems (2021)
47. Mastan, I.D., Raman, S.: Deepcfl: Deep contextual features learning from a single image. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) pp. 2896–2905 (2021)
48. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
49. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)

50. Navasardyan, S., Ohanyan, M.: Image inpainting with onion convolutions. In: Proceedings of the Asian Conference on Computer Vision (2020)
51. Park, J., Joo, K., Hu, Z., Liu, C., Kweon, I.: Non-local spatial propagation network for depth completion. In: Proc. of European Conference on Computer Vision (ECCV) (2020)
52. Peng, J., Liu, D., Xu, S., Li, H.: Generating diverse structure for image inpainting with hierarchical vq-vae. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10775–10784 (2021)
53. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
54. Rosin, P.L., Lai, Y.K., Shao, L., Liu, Y.: RGB-D Image Analysis and Processing. Springer (2019)
55. Sabini, M., Rusak, G.: Painting outside the box: Image outpainting with gans. ArXiv **abs/1808.08483** (2018)
56. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in neural information processing systems. pp. 2234–2242 (2016)
57. Schubert, S., Neubert, P., Pöschmann, J., Pretzel, P.: Circular convolutional neural networks for panoramic images and laser data. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 653–660. IEEE (2019)
58. Shan, Q., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M.: Photo uncrop. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 16–31. Springer International Publishing, Cham (2014)
59. Suin, M., Purohit, K., Rajagopalan, A.: Distillation-guided image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2481–2490 (2021)
60. Sumantri, J.S., Park, I.K.: 360 panorama synthesis from a sparse set of images on a low-power device. IEEE Transactions on Computational Imaging **6**, 1179–1193 (2020). https://doi.org/10.1109/TCI.2020.3011854
61. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: 2022 IEEE Winter Conference on Applications of Computer Vision (WACV) (2022)
62. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
63. Tang, J., Tian, F.P., Feng, W., Li, J., Tan, P.: Learning guided convolutional network for depth completion. IEEE Transactions on Image Processing **30**, 1116–1129 (2020)
64. Van Gansbeke, W., Neven, D., De Brabandere, B., Van Gool, L.: Sparse and noisy lidar completion with rgb guidance and uncertainty. In: 2019 16th International Conference on Machine Vision Applications (MVA). pp. 1–6 (2019). https://doi.org/10.23919/MVA.2019.8757939
65. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
66. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 4672–4681 (2021)

67. Wang, B., An, J.: Fis-nets: Full-image supervised networks for monocular depth estimation. CoRR **abs/2001.11092** (2020), https://arxiv.org/abs/2001.11092
68. Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In: CVPR (2020)
69. Wang, M., Lai, Y.K., Liang, Y., Martin, R.R., Hu, S.M.: Biggerpicture: Data-driven image extrapolation using graph matching. ACM Trans. Graph. **33**(6) (Nov 2014). https://doi.org/10.1145/2661229.2661278
70. Wang, N., Li, J., Zhang, L., Du, B.: Musical: Multi-scale image contextual attention learning for inpainting. In: IJCAI. pp. 3748–3754 (2019)
71. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
72. Wang, W., Zhang, J., Niu, L., Ling, H., Yang, X., Zhang, L.: Parallel multi-resolution fusion network for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14559–14568 (2021)
73. Wang, Y., Tao, X., Shen, X., Jia, J.: Wide-context semantic image extrapolation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1399–1408 (2019). https://doi.org/10.1109/CVPR.2019.00149
74. Wu, X., Li, R.L., Zhang, F.L., Liu, J.C., Wang, J., Shamir, A., Hu, S.M.: Deep portrait image completion and extrapolation. IEEE Transactions on Image Processing **29**, 2344–2355 (2020). https://doi.org/10.1109/tip.2019.2945866
75. Xie, C., Liu, S., Li, C., Cheng, M.M., Zuo, W., Liu, X., Wen, S., Ding, E.: Image inpainting with learnable bidirectional attention maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8858–8867 (2019)
76. Xu, Y., Zhu, X., Shi, J., Zhang, G., Bao, H., Li, H.: Depth completion from sparse lidar data with depth-normal constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
77. Yang, S.T., Wang, F.E., Peng, C.H., Wonka, P., Sun, M., Chu, H.K.: Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3363–3372 (2019)
78. Yang, Z., Dong, J., Liu, P., Yang, Y., Yan, S.: Very long natural scenery image prediction by outpainting. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10561–10570 (2019)
79. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018)
80. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4471–4480 (2019)
81. Zeng, W., Karaoglu, S., Gevers, T.: Joint 3d layout and depth prediction from a single indoor panorama image. In: European Conference on Computer Vision. pp. 666–682. Springer (2020)
82. Zhang, L., Wang, J., Shi, J.: Multimodal image outpainting with regularized normalized diversification. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 3422–3431 (2020). https://doi.org/10.1109/WACV45572.2020.9093636
83. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)

84. Zhang, X., Chen, F., Wang, C., Tao, M., Jiang, G.P.: Sienet: Siamese expansion network for image extrapolation. IEEE Signal Processing Letters **PP**, 1–1 (08 2020). https://doi.org/10.1109/LSP.2020.3019705
85. Zhang, Y., Funkhouser, T.: Deep depth completion of a single rgb-d image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
86. Zhang, Y., Xiao, J., Hays, J., Tan, P.: Framebreak: Dramatic image extrapolation by guided shift-maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1171–1178 (2013)
87. Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., Lu, D.: Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5741–5750 (2020)
88. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=sSjqmfsk95O
89. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1438–1447 (2019)
90. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX. Lecture Notes in Computer Science, vol. 12354, pp. 519–535. Springer (2020). https://doi.org/10.1007/978-3-030-58545-7_30, https://doi.org/10.1007/978-3-030-58545-7_30