3D-Aware Indoor Scene Synthesis with Depth Priors

Zifan Shi[1], Yujun Shen[2], Jiapeng Zhu[1], Dit-Yan Yeung[1], and Qifeng Chen[1]

[1]The Hong Kong University of Science and Technology [2]ByteDance Inc.
{vivianszf9, shenyujun0302, jengzhu0}@gmail.com
{dyyeung, cqf}@cse.ust.hk

Abstract. Despite the recent advancement of Generative Adversarial Networks (GANs) in learning 3D-aware image synthesis from 2D data, existing methods fail to model indoor scenes due to the large diversity of room layouts and the objects inside. We argue that indoor scenes do not have a shared intrinsic structure, and hence only using 2D images cannot adequately guide the model with the 3D geometry. In this work, we fill in this gap by introducing depth as a 3D prior.[1] Compared with other 3D data formats, depth better fits the convolution-based generation mechanism and is more easily accessible in practice. Specifically, we propose a dual-path generator, where one path is responsible for depth generation, whose intermediate features are injected into the other path as the condition for appearance rendering. Such a design eases the 3D-aware synthesis with explicit geometry information. Meanwhile, we introduce a switchable discriminator both to differentiate real *v.s.* fake domains and to predict the depth from a given input. In this way, the discriminator can take the spatial arrangement into account and advise the generator to learn an appropriate depth condition. Extensive experimental results suggest that our approach is capable of synthesizing indoor scenes with impressively good quality and 3D consistency, significantly outperforming state-of-the-art alternatives.[2]

Keywords: 3D-aware image synthesis, scene synthesis, depth priors

1 Introduction

Generative Adversarial Networks (GANs) [12] have enabled high-fidelity 2D image synthesis, but how to make a GAN model aware of 3D information remains unsolved. Along with the recent advent of Neural Radiance Field (NeRF) [22] for 3D scene reconstruction, some attempts [4, 5, 13, 27, 35, 41] propose to incorporate NeRF into GANs to learn a 3D-aware image generator from a 2D image collection. Instead of using 2D convolutional layers, the generator is asked to learn a point-wise implicit function, which maps the 3D coordinates to volume densities and colors [22, 35].

[1] Depth is essentially a 2.5D prior, but in this paper we use 3D for simplicity
[2] Project page can be found here.

Fig. 1. Photo-realistic 3D-aware synthesis results on bedrooms and kitchens. Left: Two sets of synthesized depth maps and their corresponding rendered images from three different viewpoints. Right: Visualization of the 3D reconstruction results (using [32] and [45]) from the synthesized samples

Although existing methods show promising results in learning 3D-aware object synthesis, such as human faces and cars, they exhibit severe performance degradation on indoor scene datasets, such as bedrooms and kitchens. There are mainly two reasons. First, objects normally have a shared intrinsic structure, which eases the difficulty of modeling 3D geometry from 2D images *only*. For instance, human heads share similar shapes, and each face consists of two eyes located at relatively defined positions. On the contrary, indoor scenes have much higher diversity, considering the complex room layout and the interior decoration [42]. Second, existing methods assume the distribution of camera poses [26, 35]. Such an assumption is sound under the case of object synthesis because objects are commonly placed at the center of a 2D image. Indoor scene images are usually shot from far more diverse viewpoints, making it too challenging for the NeRF-based approaches to handle.

In this work, we propose a new paradigm for 3D-aware image synthesis by explicitly introducing a 3D prior into 2D GANs. Compared with the volume renderer equipped with Multi-Layer Perceptron (MLP) [4, 27, 35], GANs built on Convolutional Neural Network (CNN) achieve much more appealing synthesis performances [17–19], especially from the image quality and the image resolution perspectives. Among numerous 3D data formats, such as point cloud [33, 34], voxel [2, 7], and implicit surface [21, 29], we choose depth as our prior as it is defined in the 2D domain and hence naturally suitable for the convolution-based generator. In addition, there are many publicly available depth datasets [8, 20, 23] and depth predictors [31, 43], making depth data easily accessible in practice.

To sufficiently leverage the depth prior, we re-design the objectives of both the generator and the discriminator in a conventional GAN. For one thing, we

ask the generator to synthesize a 2D image accompanied by its corresponding depth. To meet this goal, we carefully tailor a dual-path architecture, where the appearance-path takes the multi-level feature maps from the depth-path as the input conditions. Through such a design, we manage to explicitly inject the geometry information into the generator. For another, unlike the conventional discriminator that makes the real/fake decision from the 2D space, we learn a 3D-aware switchable discriminator. Specifically, it is asked to distinguish the real and synthesized samples based on the image-depth joint distribution and, simultaneously, predict the depth from an input image. The depth prediction is trained on real data and further used to supervise the fake data. In this way, the discriminator is able to gain more knowledge on the spatial layout and better guide the generator from the 3D perspective.

We evaluate our approach, termed as **DepthGAN**, on a couple of challenging indoor scene datasets. Both qualitative and quantitative results demonstrate the sufficient superiority of DepthGAN over existing methods. For example, we improve Fréchet Inception Distance (FID) [15] from 15.560 to 4.797 on the LSUN bedroom dataset [44] in 256×256 resolution. 3D visualization on a set of synthesized images is shown in Fig. 1.

2 Related Work

GAN-based Image Synthesis. With the advent of Generative Adversarial Networks (GANs) [12], a large number of works have been proposed to generate high-quality photorealistic images [3,16,18,19]. To gain explicit control of the images, researchers study the disentanglement of different properties such as poses. Supervised methods [36] leverage off-the-shelf attribute classifiers or image transformations to annotate the synthesized data and use the labeled data to guide the subspace learning in the latent space. Unsupervised methods [14,37,46] learn the control by analyzing the statistics or the model weights. While these works can control the poses with the azimuth and elevation angles, the changes may violate the consistency in the 3D space since there is no such constraint.

3D-aware Image Synthesis. Realizing that previous image synthesis methods do not consider 3D geometry, a large number of works have started to add 3D constraints for image synthesis. Voxel-based methods [24,25] learn low-dimensional 3D representations with deep voxels, followed by a learnable 3D-to-2D projection. Inspired by NeRF [22], some works [4,5,9,13,27,30,35,41] incorporate neural radiance fields for 3D-aware image generation and render more consistent images. RGBD-GAN [28] synthesizes RGBD images under two views and warps them to each other to ensure 3D consistency. In contrast, we synthesize RGB images conditioned on depth images with carefully designed architectures, which models the geometry-appearance relationship better. All the works mentioned above learn geometry and appearance from 2D RGB images alone. Due to the complexity of 3D geometry modeling and the lack of explicit 3D information, they target objects or well-aligned scenes and fail to generate high-quality 3D-aware images for complex scenes like bedrooms and kitchens. On the contrary, some other

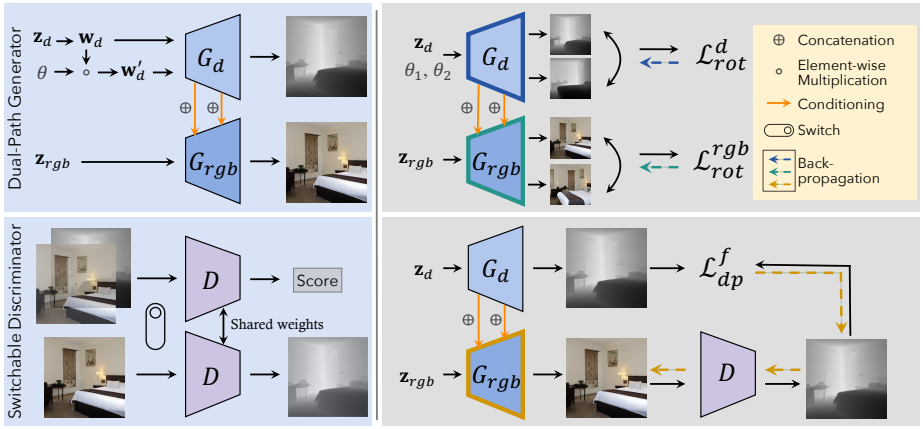

Fig. 2. Framework of DepthGAN, consisting of a *dual-path* generator that takes in two latent codes to generate the RGBD image with the appearance conditioned on the geometry, and a *switchable* discriminator that produces the realness score from an RGBD image and predicts the depth map from an RGB image. Black arrows indicate the forward computation, while dashed arrows under different colors stand for the back-propagation regarding different objective functions

works utilize 3D prior knowledge to facilitate the learning of 3D consistency. Some researchers [6,40,47] select shape as the 3D prior and use the expensive 3D-conv-based GAN to learn the geometry information, which is costly and unable to model fine details of the shape. Others [1,6] utilize more than one 3D prior, such as albedo maps and normal maps, resulting in multiple 2D GANs to learn all the 3D attributes. Instead of generating objects only, S^2-GAN [39] synthesizes indoor scenes with the help of normal maps, but it adopts the two-stage training to learn geometry first and the appearance next. All these works either have separate 3D and 2D discriminators to learn geometry and texture distributions independently, or use 2D discriminators only to make the real/fake decision on one 3D attribute or the appearance. In contrast, our discriminator is endowed with 3D and 2D knowledge simultaneously. GSN [10] follows the NeRF rendering structure and adds another depth channel in the discriminator to incorporate 3D priors, but it fails to generate images with a large diversity and reasonable fidelity due to the complex rendering process, the special requirements of training data, and the inadequacy of its discriminator.

3 Method

In this work, we propose a new paradigm for 3D-aware image synthesis via introducing depth as a 3D prior into 2D GANs. To adequately use the depth prior, we re-design both the generator and the discriminator in conventional GANs [12]. Concretely, we propose a *dual-path* generator and a *switchable* discriminator

based on the recent StyleGAN2 [19] model. The overall framework is shown in Fig. 2. For simplicity, we denote the RGB image, RGBD image, and depth image with $\mathbf{I}_{rgb}$, $\mathbf{I}_{rgbd}$, and $\mathbf{I}_d$, respectively.

3.1 Dual-Path Generator

To make the generator become aware of the geometry information, we ask it to synthesize the RGB image conditioned on the depth image. For this purpose, we tailor a dual-path generator, consisting of a depth generator G_d and an appearance renderer G_{rgb}. Two latent spaces, $\mathcal{Z}_d$ and $\mathcal{Z}_{rgb}$, are introduced to enable the independent sampling of depth and appearance. To make sure the appearance is properly rendered on top of the geometry, we feed the intermediate feature maps of G_d into G_{rgb} as the condition.

Depth Generator. To control the viewing point of the generated depth, we uniformly sample an angle θ from $[\theta_L, \theta_R]$. Since networks tend to learn better information from high-frequency signals [38], we encode θ with

$$\gamma(\theta, t) = h(sin(\theta), cos(\theta), ..., sin(t\theta), cos(t\theta)), \tag{1}$$

where t determines the maximum frequency. $h : \mathbb{R}^{2t} \rightarrow \mathbb{R}^m$ stands for a non-linear mapping, which is implemented by a two-layer fully-connection (FC) and a Leaky ReLU activation in between. Like StyleGAN2 [19], the raw depth latent code $\mathbf{z}_d \in \mathcal{Z}_d$ is projected into a more disentangled latent space, resulting in $\mathbf{w}_d \in \mathbb{R}^m$. The angle information is then injected to $\mathbf{w}_d$ through

$$\mathbf{w}'_d = \mathbf{w}_d \circ \gamma(\theta, t), \tag{2}$$

where $\circ$ denotes the element-wise multiplication. $\mathbf{w}'_d$ guides G_d on synthesizing the depth image, $\mathbf{I}^f_d$, via layer-wise style modulation [19]. Note that only the first two layers of G_d employ $\mathbf{w}'_d$ while the remaining layers still use $\mathbf{w}_d$, because only early layers correspond to the viewing point of the output image [42].

Depth-Conditioned Appearance Renderer. G_{rgb} shares a similar structure as G_d with three modifications. First, the number of output channels is 3 ($\mathbf{I}^f_{rgb}$) instead of 1 ($\mathbf{I}^f_d$). Second, G_{rgb} does not take the angle θ as the input. Third, most importantly, G_{rgb} takes the intermediate feature maps of G_d as the conditions to acquire the geometry information. Specifically, we first concatenate the per-layer feature $\mathbf{\Psi}_i$ of G_d with that of G_{rgb}, $\mathbf{\Phi}_i$. Here, i denotes the layer index. We then transform the concatenated result with

$$\mathbf{\Phi}'_i = f(\mathbf{\Psi}_i \oplus \mathbf{\Phi}_i), \tag{3}$$

where $\oplus$ stands for the concatenation operation, and f is implemented with a two-layer convolution. $\mathbf{\Phi}'_i$ has the same number of channels as $\mathbf{\Phi}_i$.

3.2 Switchable Discriminator

Unlike the discriminator in conventional GANs that simply differentiates the real and fake domains from the RGB image space, we propose a switchable

discriminator to compete with the generator by taking the spatial arrangement into account. This is achieved from two aspects. On one hand, D makes the real/fake decision based on the joint distribution of RGB images and the corresponding depths. In other words, D takes an RGBD image as the input and outputs the realness score. On the other hand, to better capture the relationship between the image and the depth, we ask D to predict the depth from a given RGB image. Concretely, we introduce a separate branch on top of some intermediate feature maps of D for depth prediction. Detailed structure of the depth prediction branch can be found in the *Supplementary Material*.

To summarize, D switches between the 4-channel RGBD inputs (*i.e.*, for realness discrimination) and the 3-channel RGB inputs (*i.e.*, for depth prediction). To achieve this goal, we come up with a switchable input layer that adaptively adjusts the number of convolutional kernels.

3.3 Training Objectives

Adversarial Loss. We adopt the standard adversarial loss for GAN training:

$$\mathcal{L}_{adv}^{d} = -\mathbb{E}[\log(D(\mathbf{I}_{rgbd}^{r}))] - \mathbb{E}[\log(1 - D(\mathbf{I}_{rgbd}^{f}))], \tag{4}$$

$$\mathcal{L}_{adv}^{g} = -\mathbb{E}[\log(D(\mathbf{I}_{rgbd}^{f}))], \tag{5}$$

where $\mathbf{I}_{rgbd}^{r}$ represents the real RGBD data, and $\mathbf{I}_{rgbd}^{f}$ concatenates the generated RGB image $\mathbf{I}_{rgb}^{f}$ and the conditioned depth $\mathbf{I}_{d}^{f}$.

Rotation Consistency Loss. We design the rotation consistency loss [28] to enhance the consistency between the synthesis from different viewpoints, *i.e.*, θ. Specifically, two angles, θ_1 and θ_2, are randomly sampled, leading to two samples, $\mathbf{I}_{rgbd,1}^{f}$ and $\mathbf{I}_{rgbd,2}^{f}$ with the same latent codes, $\mathbf{z}_d$ and $\mathbf{z}_{rgb}$. We fix the camera and rotate the scene around its central axis. We assume an underlying camera intrinsic parameter, $\mathbf{K}$, which is fixed in the training process. After rotating $\mathbf{I}_{rgbd,1}^{f}$ from θ_1 to θ_2, we will get

$$P(\mathbf{I}_{rgbd,1}^{f,rot}) = \mathbf{K}R(\theta_1, \theta_2)\mathbf{K}^{-1}P(\mathbf{I}_{rgbd,1}^{f}), \tag{6}$$

where $R(\cdot,\cdot)$ denotes the rotation operation based on the depth image, $\mathbf{I}_{d,1}^{f}$, and $P(\cdot)$ represents the coordinates of the pixels. More details are available in the *Supplementary Material*.

The rotation consistency losses $\mathcal{L}_{rot}^{d}$ and $\mathcal{L}_{rot}^{rgb}$ for the dual-path generator are then defined as

$$\mathcal{L}_{rot}^{d} = \|\mathbf{I}_{d,1}^{f,rot} - \mathbf{I}_{d,2}^{f}\|_1, \tag{7}$$

$$\mathcal{L}_{rot}^{rgb} = \|\mathbf{I}_{rgb,1}^{f,rot} - \mathbf{I}_{rgb,2}^{f}\|_1, \tag{8}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm.

Depth Prediction Loss. As discussed above, besides differentiating real and fake data, our switchable discriminator is also asked to predict the depth from a given RGB image. Such a prediction is trained on real image-depth pairs, and further used to guide the synthesis. Following [11], our depth prediction is learned with a k-class classification. Thus, the depth prediction branch of D, $D_d(\cdot)$, produces a k-channel output map, indicating the class probability for each pixel. The loss function is formulated as

$$\mathcal{L}_{dp}^{r} = \mathcal{H}(D_d(\mathbf{I}_{rgb}^{r}), \mathbf{I}_{d}^{r}), \tag{9}$$

where $\mathcal{H}(\cdot, \cdot)$ denotes the pixel-wise cross-entropy loss. $\mathbf{I}_{rgb}^{r}$ and $\mathbf{I}_{d}^{r}$ stand for the ground-truth pair.

In order to help the generated appearance, $\mathbf{I}_{rgb}^{f}$, better fit the geometry, $\mathbf{I}_{d}^{f}$, we also predict the depth from the synthesized image to in turn guide the generator with

$$\mathcal{L}_{dp}^{f} = \mathcal{H}(D_d(\mathbf{I}_{rgb}^{f}), \mathbf{I}_{d}^{f}). \tag{10}$$

Full Objectives. In summary, the dual-path generator (*i.e.*, G_d and G_{rgb}) and the switchable discriminator (*i.e.*, D) are jointly optimized with

$$\mathcal{L}_{G_d} = \mathcal{L}_{adv}^{g} + \lambda_1 \mathcal{L}_{rot}^{d}, \tag{11}$$

$$\mathcal{L}_{G_{rgb}} = \mathcal{L}_{adv}^{g} + \lambda_2 \mathcal{L}_{rot}^{rgb} + \lambda_3 \mathcal{L}_{dp}^{f}, \tag{12}$$

$$\mathcal{L}_{D} = \mathcal{L}_{adv}^{d} + \lambda_4 \mathcal{L}_{dp}^{r}, \tag{13}$$

where $\{\lambda_i\}_{i=1}^{4}$ are loss weights to balance different terms.

4 Experiments

Datasets. We conduct experiments on LSUN bedroom and kitchen datasets [44]. Details are available in the *Supplementry Material*.

Metrics. We use the following metrics for evaluation: Fréchet Inception Distance (FID) [15], Chamfer Distance (CD), Rotation Precision (RP), and Rotation Consistency (RC). FID evaluates the quality of both the generated RGB images and depth images. FID for depth images is obtained by repeating the one-channel depth image to a three-channel image as input. CD measures the 3D consistency in 3D space, which computes the cross-view distance via warping point clouds. In addition, we propose another two metrics for evaluation. *(1) Rotation Precision (RP)* is aimed to measure the accuracy of the angle of rotation given two generated images from different views of the same scene. The formulation is the same as Eq. (7), and it evaluates depths in the range [0, 1]. *(2) Rotation Consistency (RC)* targets at the rotation consistency evaluation and has the same format as Eq. (8). It evaluates RGB images with pixel range normalized to [-1, 1]. Since our discriminator can be the role of depth estimator, in ablation studies, we also report the depth prediction accuracy *DP (Real)* on real images from the test set. Besides, we predict the depth for the synthesized RGB image by the pre-trained

Table 1. Quantitative comparisons with existing 3D-aware image synthesis models on the LSUN bedroom and kitchen datasets [44] under both 128×128 and 256×256 resolutions. FID [15] regarding RGB images and depths, rotation precision (RP), rotation consistency (RC) and Chamfer distance (CD) are used as the metrics to evaluate the synthesis quality and the 3D controllability. CD is reported in the order of 10^{-3}. ↓ means lower value is better

(a) Evaluation on bedrooms

Method	128 × 128					256 × 256				
	FID↓	FID (D)↓	RP↓	RC↓	CD↓	FID↓	FID (D)↓	RP↓	RC↓	CD↓
2D-GAN SeFa [37]	8.650	-	0.572	1.027	-	7.190	-	0.401	1.110	-
RGBD-GAN [28]	28.694	370.526	-	-	-	59.026	360.858	-	-	-
GRAF [35]	63.940	184.379	0.218	1.149	44.321	66.856	188.368	0.219	0.880	66.702
GRAF(D) [35]	158.503	107.653	0.135	1.108	17.590	194.260	156.081	0.154	1.193	52.176
GIRAFFE [27]	48.412	422.634	-	-	-	44.232	420.681	-	-	-
π-GAN [4]	28.128	201.722	0.033	0.572	0.744	48.926	174.744	0.052	0.597	3.403
π-GAN(D) [4]	30.932	101.739	**0.022**	**0.420**	**0.355**	49.640	94.196	0.036	0.510	1.201
StyleNeRF [13]	13.675	284.088	0.140	1.026	0.841	15.560	288.379	0.159	1.055	2.352
VolumeGAN [41]	18.121	175.963	0.088	0.707	1.599	17.345	164.190	0.110	0.672	4.038
DepthGAN (Ours)	**4.040**	**18.874**	0.040	0.530	0.461	**4.797**	**17.140**	**0.025**	**0.456**	**0.339**

(b) Evaluation on kitchens

Method	128 × 128					256 × 256				
	FID↓	FID (D)↓	RP↓	RC↓	CD↓	FID↓	FID (D)↓	RP↓	RC↓	CD↓
2D-GAN SeFa [37]	11.530	-	0.748	1.115	-	10.850	-	0.480	1.163	-
RGBD-GAN [28]	33.425	267.036	-	-	-	51.044	392.126	-	-	-
GRAF [35]	86.920	239.657	0.224	1.326	48.265	94.095	204.050	0.227	0.928	72.472
GRAF(D) [35]	139.902	157.801	0.110	0.970	19.237	244.480	142.436	0.133	0.966	52.775
GIRAFFE [27]	42.923	307.233	-	-	-	50.256	370.760	-	-	-
π-GAN [4]	29.790	398.146	0.028	0.702	0.832	41.178	398.946	0.051	0.726	3.833
π-GAN(D) [4]	46.332	112.171	**0.025**	**0.482**	**0.258**	77.066	104.865	0.039	0.566	1.092
DepthGAN (Ours)	**5.068**	**17.655**	0.038	0.551	0.468	**6.051**	**25.335**	**0.028**	**0.502**	**0.329**

depth prediction model [43], and compare it with the generated depth to form the evaluation metric *DP (Fake)*.

Baselines.[3] We compare against seven state-of-the-art methods for 3D-aware image synthesis: HoloGAN [24], RGBD-GAN [28], GRAF [35], GIRAFFE [27], π-GAN [4], StyleNeRF [13] and VolumeGAN [41]. For a fair comparison, we also incorporate depth information into existing methods by either employing another discriminator for depth learning or changing the input/output from RGB to RGBD image, which result in the two variants, GRAF(D) and π-GAN(D). Another baseline is a 2D-based approach named SeFa [37], which can rotate the scenes through interpolation in the latent space.

4.1 Quantitative Results

Tab. 1 reports the quantitative comparisons. Since RGBD-GAN does not learn geometry at all as shown in Fig. 3 and Fig. 4, it makes no sense to evaluate it with RP, RC, and CD. GIRAFFE fixes the background and rotates objects only, and thus we do not report RP, RC, and CD on it. We show significant improvement

[3] More details of the baselines can be found in the *Supplementary Material*.

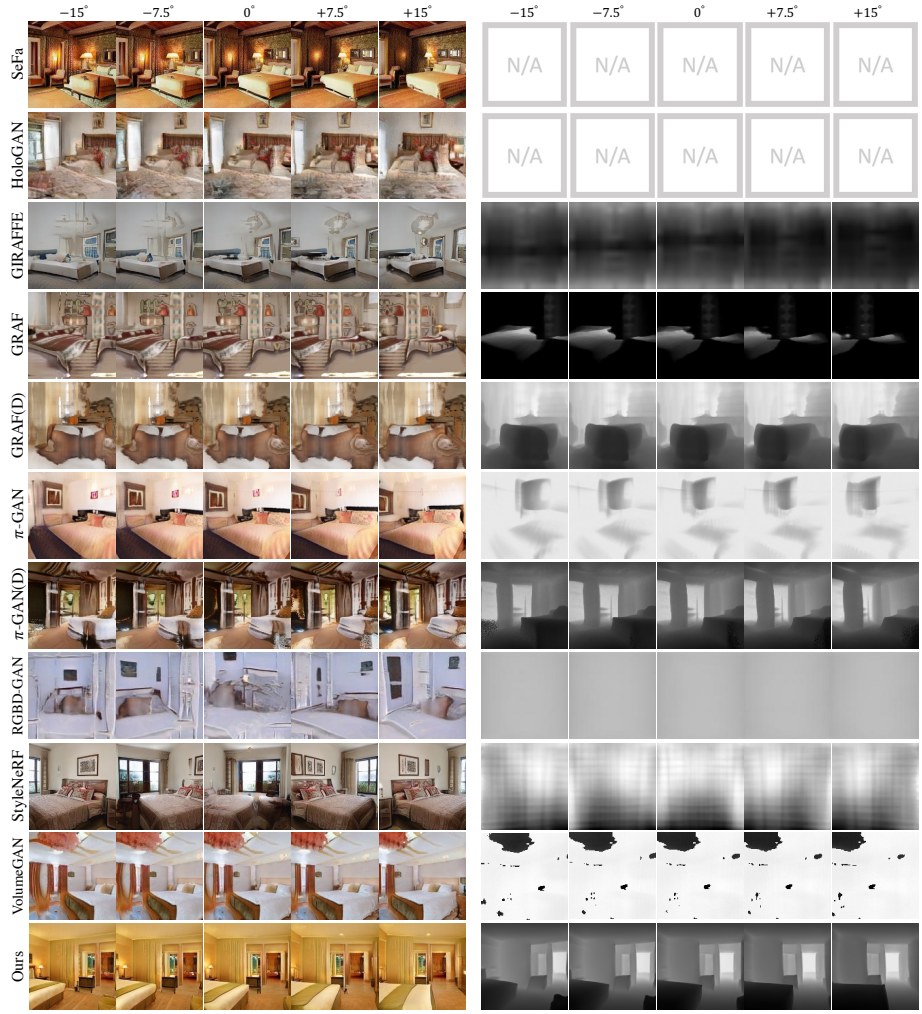

Fig. 3. Qualitative comparisons on LSUN bedroom [44] with existing 3D-aware image synthesis models. *Left:* RGB images. *Right:* The corresponding depths. Each scene is evenly rotated by 30 degrees to generate five samples. Zoom in for details

of image quality compared with 3D-aware image synthesis baselines in terms of FID scores on both RGB images and depth images. When 3D-aware image synthesis methods are given with the depth information for training, the quality of the geometry generally improves, but the quality of the appearance decreases. While maintaining the high quality of images, ours ensures the 3D consistency as well. Note that while π-GAN has lower RP, RC, and CD values than ours, it produces depth maps with simple geometry reflected by FID on depth images and the qualitative results, which makes it easier to maintain consistency.

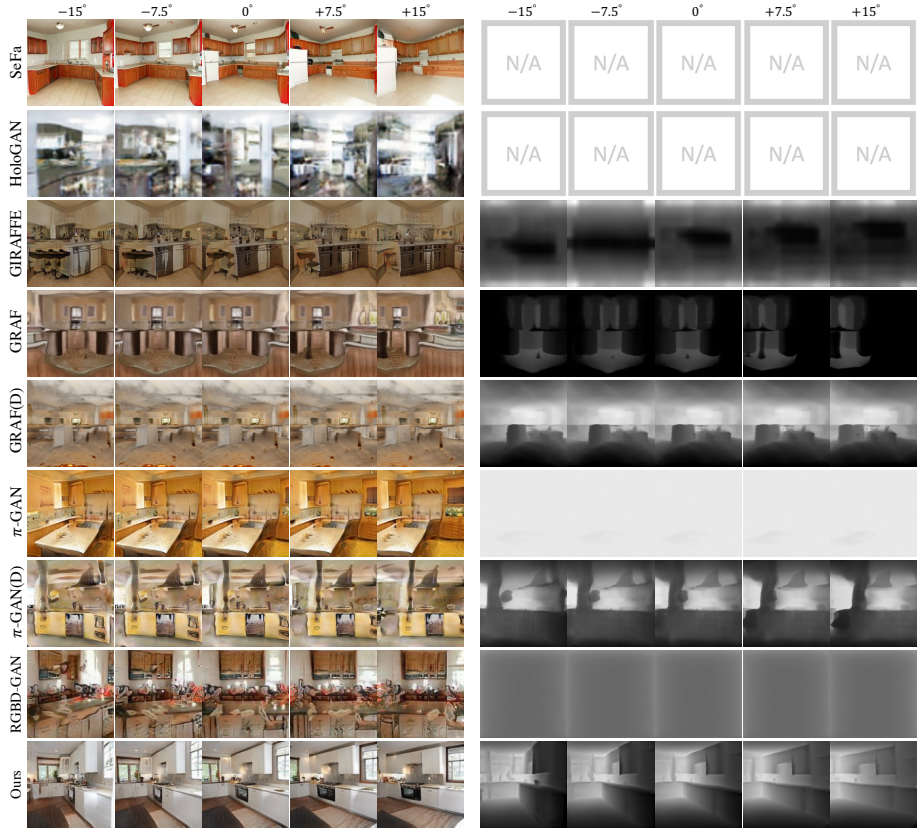

Fig. 4. Qualitative comparisons on LSUN kitchen [44] with existing 3D-aware image synthesis models. *Left:* RGB images. *Right:* The corresponding depths. Each scene is evenly rotated by 30 degrees to generate five samples. Zoom in for details

4.2 Qualitative Results

The generated images from each baseline and our DepthGAN are shown in Fig. 3 and Fig. 4. 2D-GANs can generate RGB images of high quality. However, interpolation in the latent space does not guarantee 3D consistency, and thus both the geometry and appearance can be changed during rotation. Though 3D-aware image synthesis methods can synthesize RGB images of discernible scenes, they generally fail to learn a reasonable geometry unsupervisedly for both the bedrooms and the kitchens. This indicates that in the previous works, generating a visually-pleasing RGB image does not require a good understanding of the underlying 3D geometry. Besides, the image quality degrades significantly compared with that of 2D GANs. With the help of ground-truth depth information, GRAF(D) and π-GAN(D) can generate geometries of higher quality but sacrifice the quality of the appearance. In contrast, our DepthGAN can generate images with reasonable

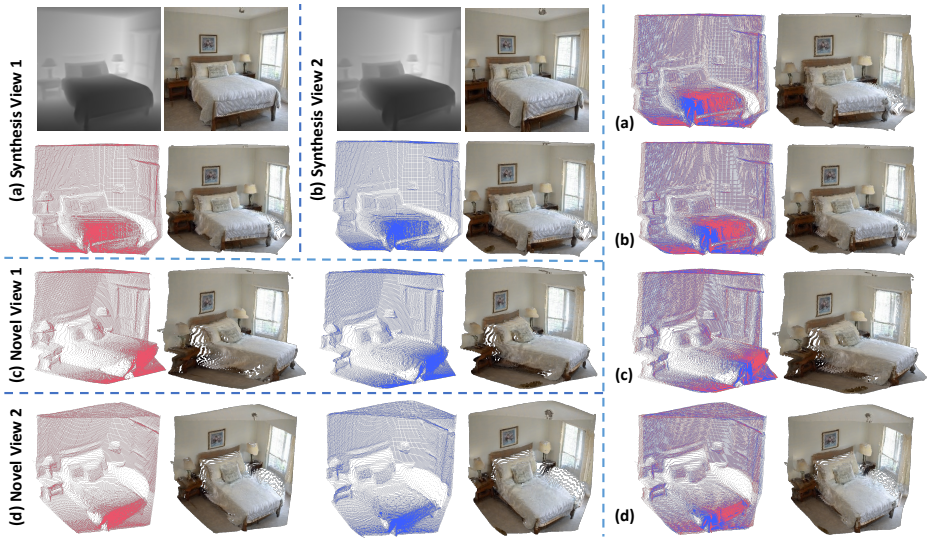

Fig. 5. Geometry visualization. ***Left:*** (a, b) Original syntheses (*i.e.*, depth and RGB on the top row) from two different views by our dual-path generator, as well as the extracted geometry (*i.e.*, point cloud and rendered appearance on the bottom row). (c, d) Two novel views through rotating (a, b). ***Right:*** Overlaying the two synthesized geometry and visualizing them from four different views (same as those on the *Left*). Zoom in for details

geometries and photo-realistic appearance simultaneously, which mitigates the gap between the 3D-aware image synthesis and 2D GANs and surpass the recent 3D-aware image synthesis methods on scene generation as well.

We also analyze the consistency by visualizing the geometry learnt by DepthGAN in Fig. 5, where we drag point clouds from two RGBD images to the same view and overlay them. The overlapped point clouds on the right side demonstrate geometry rendered from two views are consistent with each other (see pillow, lamp, and edge of bed). Besides, they complement each other for a more complete point cloud (*i.e.*, fewer holes on the overlapped point clouds).

4.3 Ablation Study

We analyze the effectiveness of each component of DepthGAN. Evaluation results are shown in Tab. 2. There is a discrepancy between the generated depth and RGB images if there is no depth prediction loss. As such, the discriminator struggles to lead the generator to capture a coherent relationship between the geometry and the appearance, and all the metrics drop significantly. Without the rotation-consistency loss on RGB images, the RGB consistency completely depends on the conditioning and the discriminator, which forces the network to figure out the consistency on RGB images by itself. While the network is working hard to learn such consistency, it hinders other aspects of learning to some extent.

Table 2. Ablation study conducted on LSUN bedroom dataset [44] under 128×128 resolution. FID [15] regarding RGB images and depths, rotation precision (RP) and rotation consistency (RC), depth prediction (DP) on real and fake samples are used as the metrics to evaluate the synthesis quality and the 3D controllability. ↓ means lower value is better

	FID↓	FID (D)↓	RP↓	RC↓	DP (Real)↓	DP (Fake)↓
w/o L_{dp}^{r}, L_{dp}^{f}	4.882	26.518	0.067	0.711	-	0.317
w/o L_{dp}^{f}	5.441	24.633	0.066	0.683	1.334	0.318
w/o L_{rgb}^{rot}	5.038	24.834	0.067	0.716	1.303	0.315
w/o L_{d}^{rot}	4.504	**8.315**	0.152	1.196	1.279	0.311
w/o condition	24.062	119.917	0.097	1.443	1.242	0.343
w/o condition, rotation	**2.793**	21.225	-	-	1.205	0.312
Ours-full	4.040	18.874	**0.040**	**0.530**	**1.201**	**0.310**

Fig. 6. Diverse synthesis via varying the appearance latent code $\mathbf{z}_{rgb}$, with the depth latent code $\mathbf{z}_d$ fixed. We can tell that all samples are with the same geometry, benefiting from our *dual-path* generator that conditions the appearance generation branch on the depth branch

To test the performance of DepthGAN without rotation-consistency loss on depth, we allow the rotation-consistency loss on RGB images to back-propagate the gradients to G_d, which is different from our original design. Without the consistency loss on the rotation of depth images, there are fewer constraints on the depth generation, resulting in a lower FID score on the generated depth images. However, the rotation precision and consistency measurements experience a significant drop due to the lack of explicit supervision on the depth rotation. We also report the result without conditioning appearance features on depth features. For discriminator, the depth prediction from a real image is preserved to enhance the 3D knowledge within it. When rotation consistency loss is included for training, where the generator has the same structure as RGBD-GAN [28], the network is unable to capture the correct depth-appearance pair. If rotation consistency loss is removed, the generator is the same with that of StyleGAN2 [19]. Although the FID score on RGB images is lower, the network fails to view the scene from different angles directly and thus lacks 3D knowledge.

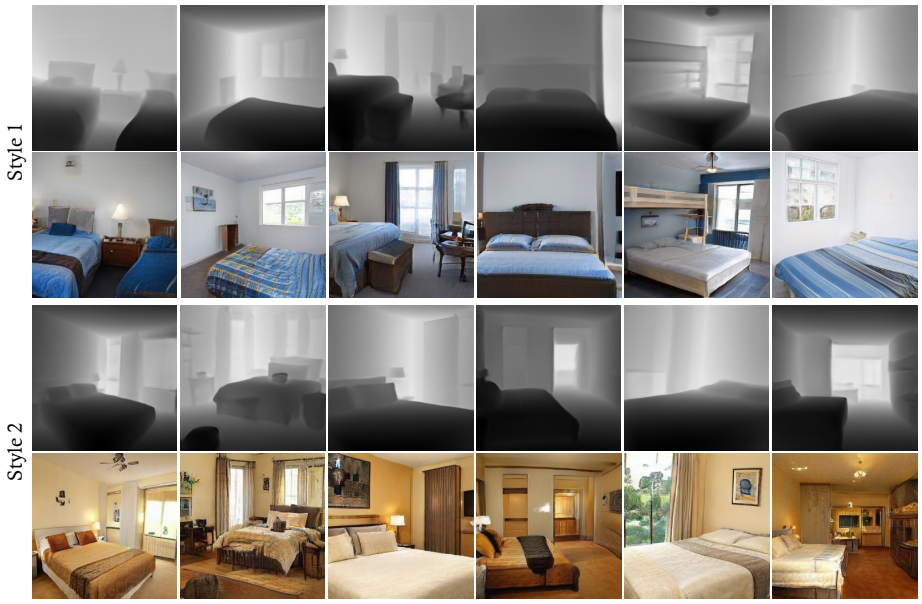

Fig. 7. Diverse geometries via varying the depth latent code $\mathbf{z}_d$, with the appearance latent code $\mathbf{z}_{rgb}$ fixed. We can observe the appropriate alignment between the depth image and the corresponding appearance, where all RGB images are rendered with the same style

4.4 Controllable Image Synthesis

Disentanglement. With the design of the dual-path generator, the latent spaces of the two generators are separate and thus can be sampled independently. This allows for clear disentanglement of the geometry and appearance. Fig. 6 shows the cases where the latent codes for depth generation are fixed, and the latent codes are changed for varying appearance. The underlying 3D geometries are the same for all the images within the same row while the styles keep changing. On the contrary, images in Fig. 7 share the same style but the geometries are various, which is brought by a fixed latent code for G_{rgb} and different latent codes for G_d.
Linearity. To demonstrate that two latent spaces learned by DepthGAN are semantically meaningful, we linearly interpolate between two latent codes from one latent space and fix the latent code from the other latent space. The interpolation results are shown in Fig. 8.

4.5 Discussion

Rotation. Although the choice of rotation axis as the central one relieves the constraint that the camera stays in the same sphere with all scenes located on the center, it brings a large variety of rotation distributions. Thus, during the rotation, the newly generated view may be out of the manifold learned during

Fig. 8. Interpolation results regarding both the depth (*i.e.*, the top two rows) and the appearance (*i.e.*, the bottom row). It it noteworthy that interpolation in the latent space is *different* from rotating the viewpoint, as the two depth codes for interpolation are not guaranteed to represent the same geometry. Instead, this figure only verifies the continuity of the latent spaces in our learned model

training, resulting in unsatisfactory images. The access to the prior distribution of the rotation axes from real data may ease the problem. As the current rotation range is $[-15°, 15°]$, we do not take special treatment for occlusion. However, it is an inevitable problem if the angle range is required to be larger, which we leave for future exploration.

Ground-Truth 3D Information. The quality of the generated 3D-aware images highly relies on the performance of the pre-trained depth prediction methods. We notice that for some objects such as light on the ceiling and some windows or paintings on the wall, there is no depth information available (e.g., images in Fig. 6 and Fig. 7). This is due to the fact that the pre-trained depth prediction model fails to predict depths for such minute details. Introducing real depth images collected by depth sensors into the training should alleviate this limitation.

5 Conclusion

In this work, we present DepthGAN, which can learn the appearance and the underlying geometry of indoor scenes simultaneously. DepthGAN takes depth as the 3D prior to facilitate the learning of 3D-aware image synthesis. A dual-path generator and a switchable discriminator are carefully designed to make sufficient use of the depth prior. Experimental results demonstrate the superiority of our approach over existing methods from both the image quality and the 3D controllability perspectives.

Acknowledgement. We thank Yinghao Xu and Sida Peng for their fruitful discussions and valuable comments.

References

1. Alhaija, H.A., Mustikovela, S.K., Geiger, A., Rother, C.: Geometric image synthesis. In: Asian Conf. Comput. Vis. (2018)
2. Ashburner, J., Friston, K.J.: Voxel-based morphometry—the methods. Neuroimage (2000)
3. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: Int. Conf. Learn. Represent. (2019)
4. Chan, E., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-GAN: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
5. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. arXiv preprint arXiv:2112.07945 (2021)
6. Chen, X., Cohen-Or, D., Chen, B., Mitra, N.J.: Towards a neural graphics pipeline for controllable image generation. Computer Graphics Forum (2021)
7. Cheung, G.K., Kanade, T., Bouguet, J.Y., Holler, M.: A real time system for robust 3d voxel reconstruction of human motions. In: IEEE Conf. Comput. Vis. Pattern Recog. (2000)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
9. Deng, Y., Yang, J., Xiang, J., Tong, X.: Gram: Generative radiance manifolds for 3d-aware image generation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
10. DeVries, T., Bautista, M.A., Srivastava, N., Taylor, G.W., Susskind, J.M.: Unconstrained scene generation with locally conditioned radiance fields. In: Int. Conf. Comput. Vis. (2021)
11. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Adv. Neural Inform. Process. Syst. (2014)
13. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In: Int. Conf. Learn. Represent. (2022)
14. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: GANSpace: Discovering interpretable GAN controls. In: Adv. Neural Inform. Process. Syst. (2020)
15. 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: Adv. Neural Inform. Process. Syst. (2017)
16. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: Int. Conf. Learn. Represent. (2018)
17. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: Adv. Neural Inform. Process. Syst. (2021)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019)
19. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)

20. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: IEEE Conf. Comput. Vis. Pattern Recog. (2015)
21. Michalkiewicz, M., Pontes, J.K., Jack, D., Baktashmotlagh, M., Eriksson, A.: Implicit surface representations as layers in neural networks. In: Int. Conf. Comput. Vis. (2019)
22. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Eur. Conf. Comput. Vis. (2020)
23. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Eur. Conf. Comput. Vis. (2012)
24. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: HoloGAN: Unsupervised learning of 3d representations from natural images. In: Int. Conf. Comput. Vis. (2019)
25. Nguyen-Phuoc, T., Richardt, C., Mai, L., Yang, Y.L., Mitra, N.: BlockGAN: Learning 3d object-aware scene representations from unlabelled images. In: Adv. Neural Inform. Process. Syst. (Nov 2020)
26. Niemeyer, M., Geiger, A.: Campari: Camera-aware decomposed generative neural radiance fields. arXiv preprint arXiv:2103.17269 (2021)
27. Niemeyer, M., Geiger, A.: GIRAFFE: Representing scenes as compositional generative neural feature fields. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
28. Noguchi, A., Harada, T.: RGBD-GAN: Unsupervised 3d representation learning from natural image datasets via rgbd image synthesis. In: Int. Conf. Learn. Represent. (2020)
29. Ohtake, Y., Belyaev, A., Seidel, H.P.: Ridge-valley lines on meshes via implicit surface fitting. In: ACM SIGGRAPH (2004)
30. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: Stylesdf: High-resolution 3d-consistent image and geometry generation. arXiv preprint arXiv:2112.11427 (2021)
31. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
32. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Int. Conf. on 3-D digital imaging and modeling (2001)
33. Rusu, R.B., Cousins, S.: 3d is here: Point cloud library (pcl). In: IEEE Int. Conf. on Robotics and Automation (2011)
34. Schnabel, R., Wahl, R., Klein, R.: Efficient ransac for point-cloud shape detection. In: Comput. Graph. Forum (2007)
35. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: GRAF: Generative radiance fields for 3d-aware image synthesis. In: Adv. Neural Inform. Process. Syst. (2020)
36. Shen, Y., Yang, C., Tang, X., Zhou, B.: InterFaceGAN: Interpreting the disentangled face representation learned by GANs. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
37. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in GANs. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
38. Sitzmann, V., Martel, J.N., Bergman, A.W., Lindell, D.B., Wetzstein, G.: Implicit neural representation with periodic activation functions. In: Adv. Neural Inform. Process. Syst. (2020)
39. Wang, X., Gupta, A.: Generative image modeling using style and structure adversarial networks. In: Eur. Conf. Comput. Vis. (2016)

40. Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B.: Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Adv. Neural Inform. Process. Syst. (2016)
41. Xu, Y., Peng, S., Yang, C., Shen, Y., Zhou, B.: 3d-aware image synthesis via learning structural and textural representations. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
42. Yang, C., Shen, Y., Zhou, B.: Semantic hierarchy emerges in deep generative representations for scene synthesis. Int. J. Comput. Vis. (2020)
43. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
44. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
45. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: A modern library for 3D data processing. arXiv:1801.09847 (2018)
46. Zhu, J., Feng, R., Shen, Y., Zhao, D., Zha, Z., Zhou, J., Chen, Q.: Low-rank subspaces in gans. In: Adv. Neural Inform. Process. Syst. (2021)
47. Zhu, J.Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J.B., Freeman, W.T.: Visual object networks: Image generation with disentangled 3D representations. In: Adv. Neural Inform. Process. Syst. (2018)