Image-to-Voxel Model Translation with Conditional Adversarial Networks

Vladimir A. Knyaz[1,2](✉), Vladimir V. Kniaz[1,2], and Fabio Remondino[3]

[1] State Research Institute of Aviation Systems (GosNIIAS), Moscow, Russia
{knyaz,vl.kniaz}@gosniias.ru
[2] Moscow Institute of Physics and Technology (MIPT), Dolgoprudny, Russia
[3] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
remondino@fbk.eu

Abstract. We present a single-view voxel model prediction method that uses generative adversarial networks. Our method utilizes correspondences between 2D silhouettes and slices of a camera frustum to predict a voxel model of a scene with multiple object instances. We exploit pyramid shaped voxel and a generator network with skip connections between 2D and 3D feature maps. We collected two datasets VoxelCity and VoxelHome to train our framework with 36,416 images of 28 scenes with ground-truth 3D models, depth maps, and 6D object poses. We made the datasets publicly available (http://www.zefirus.org/Z_GAN). We evaluate our framework on 3D shape datasets to show that it delivers robust 3D scene reconstruction results that compete with and surpass state-of-the-art in a scene reconstruction with multiple non-rigid objects.

Keywords: Conditional GAN · Voxel model · 6D pose estimation

1 Introduction

Does a voxel model with a shape $128 \times 128 \times 1$ provide any information about a 3D object? If the XY plane of the voxel model is normal to a camera optical axis, the voxel model is similar to the object's semantic segmentation.

Modern methods [27,34] demonstrate the state-of-art results on the task of semantic segmentation. Although deep networks trained for segmentation provide the resolution of an input color image, the resolution of a voxel model output produced by modern networks is lower than the resolution of an input image [26,53,58,59,61].

We hypothesize that a pixel correspondence between an input color image and slices of a voxel model can improve the quality of fine details in a voxel output. The necessary correspondences are found using three interconnected steps: (1) we provide an aligned voxel model for each color image, (2) we use

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11129, pp. 601–618, 2019.
https://doi.org/10.1007/978-3-030-11009-3_37

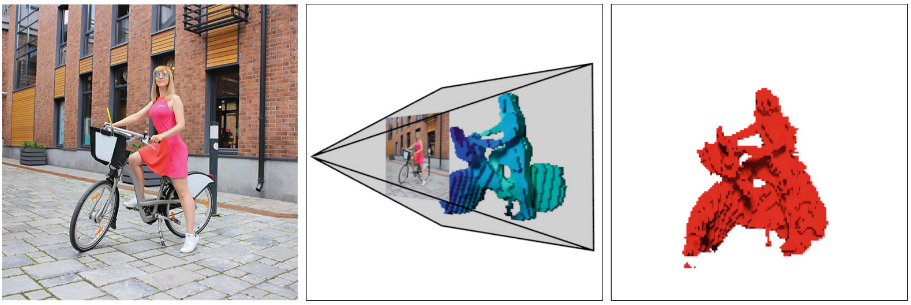


Fig. 1. Results of our image-to-voxel translation based on generative adversarial network (GAN) and frustum voxel model. Input color image (left). Ground truth frustum voxel model slices colored as a depth map (middle). The voxel model output (right). (Color figure online)

slices of a camera's frustum to built the voxel space, (3) we use a generator network with skip connections [27,46] to feed high-resolution image features to 3D deconvolutional layers of a generator network (Fig. 1).

It is challenging to predict a voxel model from a single color image. The color-to-voxel model translation problem has received a lot of scholar attention in recent time [9,17,26,44,48,53,58,59,61]. The trained models have demonstrated state-of-the-art results on large datasets with 3D object annotations [33,60]. Main limitations of the existing models are a single object focused prediction and limited generalization ability.

A research project has recently been started by the authors. The project is focused on the development of a low-cost driver assistance system. We developed two new 3D shape datasets VoxelCity and VoxelHome to train our framework. The datasets include 36,416 images of 28 scenes with ground-truth 3D models, depth maps, and 6D object poses.

The results of the trained `Z-GAN` model are encouraging. We experimented with high-resolution voxel outputs of 128 × 128 × 128 and were able to predict the shape of multiple objects accurately. We evaluated our `Z-GAN` model using the Pascal 3D+ [60] and the IKEA [33] datasets. The comparison with the state-of-the-art has demonstrated that the `Z-GAN` effectively outperforms modern models in the number of reconstructed objects, the generalization ability, and the resolution of the output voxel model. The `Z-GAN` model can be used in the 3D vision applications such as robot vision, 6D pose estimation, and 3D model reconstruction.

The rest of the paper is organized as follows. Section 2 outlines modern approaches to voxel model reconstruction. In Sect. 3 we describe the structure of our VoxelCity and VoxelHome datasets. In Sect. 4 the developed conditional `Z-GAN` model is presented. Section 5 presents the evaluation of baselines and the developed model.

1.1 Contributions

The key contributions of this paper are: (1) the conditional adversarial volumetric `Z-GAN` framework for the generation of a voxel model from a single-view color image, (2) VoxelCity and VoxelHome datasets with 36,416 color images, ground-truth voxel models, depth maps and camera orientations of 21 outdoor and 7 indoor scenes, (3) an evaluation of baselines and our framework on 3D shape datasets.

2 Related Work

Generative Adversarial Networks. Recently proposed Generative Adversarial Networks (GANs) [18] provide a mapping from a random noise vector to a domain of the desired outputs (e.g., images, voxel models). GANs are gaining increasing attention in recent years. They provide encouraging results in tasks like image-to-image translation [27] and voxel model generation [59].

Single-Photo 3D Model Reconstruction. Accurate 3D reconstruction is challenging if only a single color image is provided. This problem was always of great interest for the research community [12,42,43] and in the last years many new approaches were proposed based on the use of deep learning [9,17,26,44,48,53,58,59,61]. While a number of methods were proposed for prediction of unobserved voxels from a single depth map [15,51,62–64], prediction of the voxel model of a complex scene from a single color (RGB) image is more ambiguous. Prior knowledge of 3D shape is required for the robust performance of a single-image method. Hence, most of the methods split the problem into two steps: object recognition and a 3D shape reconstruction. In [17] a deep learning method for a single image voxel model reconstruction was proposed. The method leverages an auto-encoder architecture for a voxel model prediction. While the model has demonstrated promising results, the resolution of the voxel model was limited to $20 \times 20 \times 20$ elements. An approach that combines single-view and multi-view reconstruction modes was proposed in [9]. In [44] a new voxel decoder architecture was proposed that leverages voxel tube and shape layers to increase the resulting voxel model resolution. A comparison of surface-based and volumetric 3D model prediction is performed in [48].

3D shape synthesis from a latent space has received a lot of scholar attention recently [7,17,59]. Wu et al. have proposed a GAN model [59] for a voxel model generation (`3D-GAN`). The model was capable to predict voxel models with resolution $64 \times 64 \times 64$ from a randomly sampled noise vector. `3D-GAN` was used for a single-image 3D reconstruction using an approach proposed in [17]. While 3D models produced by the `3D-GAN` model provided more details compared to [17], the generalization ability of the approach was insufficient to predict voxel models of previously unseen 3D shapes.

3D Shape Datasets. Multiple 3D shape datasets were designed [8,33,52,60] for training deep models. Manual annotation was performed for the Pascal VOC dataset [14] to align a set of CAD models with color photos. The extended dataset was termed Pascal 3D+ [60]. While many models were trained using the Pascal 3D+ dataset, it provides a coarse correspondence between a 3D model and a photo. A large ShapeNet dataset [8] was collected to address the problems of shape recognition and generative modeling. However, training for single photo 3D model reconstruction is possible only with synthetic data. Hinterstoisser et al. have designed a large Linemod dataset [20] with aligned RGB-D data. The dataset is focused on object recognition in the indoor setting. The Linemod dataset was intensively used for training 6D pose estimation algorithms [2,3,5,6,10,22,23,32,35,40,50,55]. In [21] a large dataset for 6D pose estimation of texture-less objects was developed. An MVTec ITODD dataset [11] addresses the challenging problem of 6D pose prediction in industrial application.

The 6D pose estimation has received a lot of scholar attention recently [2,3,5,6,10,22,23,29,31,32,35,40,50,55]. Accurate estimation of camera pose relative to an object is of primary importance in such fields as an autonomous driving [1,4,36,39] and Simultaneous Localization and Mapping (SLAM) [13,57]. However, most of the datasets contain 3D data as LIDAR range scans. As no complete 3D shapes are provided in the existing datasets, they require an additional annotation for single-photo 3D reconstruction.

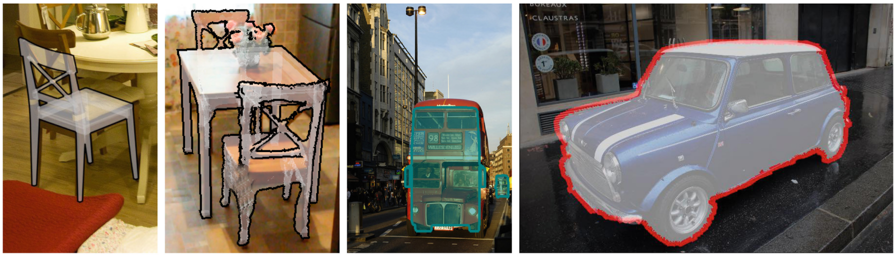
3 Dataset

We collected two new datasets VoxelCity and VoxelHome to train our `Z-GAN` model. The primary motivation for the creation of new datasets was an absence of large 3D shape datasets with pixel-level 3D object annotations. Annotations provided in Pascal 3D+ dataset [60] present CAD models of abstract classes that do not provide real silhouettes of 3D objects.

We capture multi-view images of scenes to generate our datasets (Sects. 3.2 and 3.3), composed of images, depth maps, reconstructed 3D models and ground-truth 3D CAD models. We recover 6D camera poses for each image and textured 3D models using state-of-the-art SfM algorithms [30,38,41,54]. Then we manually annotated all objects in a scene to provide multi-object 6D poses. The SfM-based approach provides two benefits. Firstly, SfM models present real configuration of objects in a scene that provides a pixel-level correspondence between images and a 3D model (Fig. 2). Secondly, SfM provides a 6D camera pose for each image. We made the datasets compliant with the SIXD Challenge dataset format [24].

3.1 3D Model Generation Using SfM

The multi-view image-based 3D reconstruction pipeline, generally called Structure from Motion (SfM), based on the integration of photogrammetric and computer vision algorithms, has become in the last years a powerful and valuable



(a) (b) (c) (d)

Fig. 2. Comparison of an alignment of 3D models in the Pix3D (a) and our VoxelHome (b) datasets and in the Pascal 3D+ (c) and our VoxelCity (d) datasets. Please note that (a) and (c) do not provide perfect alignment of the contours.

approach for 3D modeling purposes. It generally ensures sufficient automation, low cost, efficient results and ease of use, even for non-expert users. SfM now successfully reconstructs scenes containing hundred thousands or even millions of images [19,47]. Available online reconstruction services decoupled the user from a powerful hardware that carried out the reconstructions, only requiring to upload the images on a Cloud server [54]. Recently, online SfM methods demonstrated that it is possible to add new images to existing 3D reconstructions and build an incremental surface model [25,38]. We manually created 3D CAD models using the coarse 3D models reconstructed using SfM as a baseline. We use the CAD models to generate the voxel models for network training.

3.2 VoxelCity

Our VoxelCity dataset includes 3D models of 21 scenes, composed of 18,836 color images with reconstructed 3D models, ground-truth 3D CAD models, depth maps and 6D poses of seven object classes: human, car, bicycle, truck, van. Examples of 3D scenes and object pose annotations for various object classes are presented in Fig. 3. Comparison to previous 3D shape datasets regarding outdoor scenes is presented in Table 1.

Table 1. Comparison to previous outdoor 3D shape dataset. The type of data provided is listed: dense (D), coarse (C).

Dataset	#scene	#image	#class	#3D model	6D pose	P. cloud	Depth
KITTI [16]	×	15,001	3	×	✓	×	×
Pascal 3D+ [60]	×	30,364	12	77	✓	C	C
CL [28]	6	12,000	×	×	✓	D	×
VoxelCity	21	18,836	7	38	✓	D	D

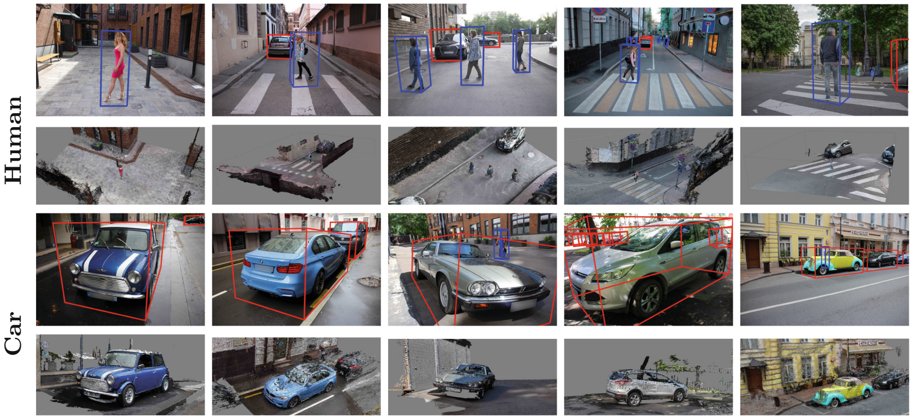


Fig. 3. Examples of color images with 6D pose annotations and ground truth dense point clouds from our VoxelCity dataset. (Color figure online)

3.3 VoxelHome

Our VoxelHome dataset presents 3D models of 7 indoor scenes, composed of 17,580 color images with reconstructed 3D models, ground-truth 3D CAD models, depth maps and 6D poses of nine object classes: chair, table, armchair, sofa, stool, cupboard, vase, washing machine, oven. Examples of 3D models and object pose annotations for various object classes are presented in Fig. 4. We present comparison to previous datasets regarding indoor scenes in Table 2.

Table 2. Comparison to previous outdoor 3D shape dataset. The type of data provided is listed: dense (D), coarse (C).

Dataset	#scene	#image	#class	#3D model	6D pose	P. cloud	Depth
IKEA [33]	×	759	7	219	✓	×	×
Linemod [20]	1	18,241	15	15	✓	×	×
T-LESS [21]	20	$3 \times 49,000$	26	30	✓	D	D
7 scenes [49]	7	43,000	×	×	✓	D	D
12 scenes [56]	12	246,673	×	×	✓	C	C
VoxelHome	7	17,580	9	64	✓	D	D

4 Method

The aim of the present research is to apply conditional generative adversarial network to the color image-to-voxel model translation task. The straightforward

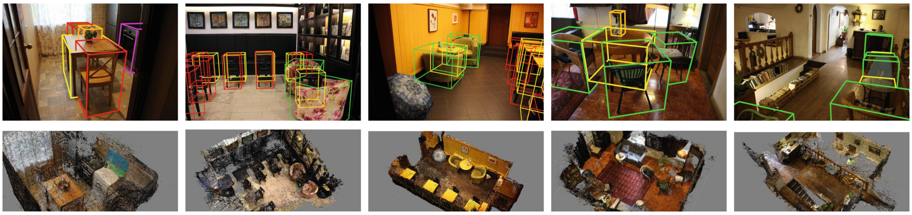


Fig. 4. Examples of color images with 6D pose annotations and ground truth dense point clouds from our VoxelHome dataset.

approach is to change the network output from an image to a voxel model. However, the convergence of the training process is poor for such setting. We hypothesize that the performance can be improved if the voxel model will be aligned with the input image.

A depth map is an example of an aligned 3D representation of the color image. While the depth map provides the 3D shape only for the visible surface of objects, the voxel model encodes the complete 3D model of the scene. We use assumptions made by [58] as the starting point for our 3D model representation. To provide the aligned voxel model, we combine depth map representation with a voxel grid. We term the resulting 3D model as a Frustum Voxel model (Fruxel model).

4.1 Frustum Voxel Model

The main idea of the fruxel model is to provide precise alignment of voxel slices with contours of a color image. Such alignment can be achieved with a common voxel model if the camera has an orthographic projection and its optical axis coincides with the Z-axis of the voxel model (see Fig. 5, left). We generalize such alignment to the perspective projection. As the camera frustum is no longer corresponding to the cube voxel elements, we use sections of a pyramid.

Fruxel model representation provides multiple advantages. Firstly, each XY slice of the model is aligned with some contours on a corresponding color photo (some parts of them can be invisible). Secondly, a fruxel model encodes a shape of both visible and invisible surfaces. Hence, unlike the depth map, it contains complete information about the 3D shapes. In other words, the fruxel model is similar to theatre scenery composed of flat screens with drawings of objects that imitate perspective space. Please note, that while fruxel elements have different dimensions in object space, all slices of the fruxel model have the same number of fruxel elements (e.g., $128 \times 128 \times 1$).

A fruxel model is characterized by a following set of parameters: $\{z_n, z_f, d, \alpha\}$, where z_n is a distance to a near clipping plane, z_f is a distance to a far clipping plane, d is the number of frustum slices, α is a field of view of a camera.

While fruxel model provides contour correspondence with a color image, its interpretation by a human may be complicated. We consider fruxel models as a

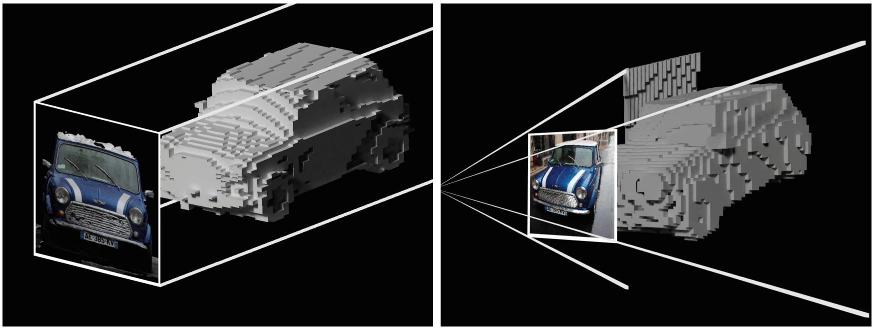


Fig. 5. Comparison between voxel model (left) and the proposed frustum voxel model (right) with shape 64 × 64 × 64 fruxel elements.

special representation of a voxel model optimized for the training of conditional adversarial networks. Nevertheless, a fruxel model can be converted into three common data types: (1) voxel model, (2) depth map, (3) object annotation.

A voxel model can be produced from the fruxel model by scaling each consequent layer slice by the coefficient k defined as follows:

$$k = \frac{z_n}{z_n + s_z}, \tag{1}$$

where $s_z = \frac{z_f - z_n}{d}$ is the size of the fruxel element along the Z-axis.

To generate a depth map P from the fruxel model, we multiply indices of the frontmost non-empty elements by the step s_z.

$$P(x, y) = \mathrm{argmax}_i [F_i(x, y) = 1] \cdot s_z \tag{2}$$

where $P(x, y)$ is an element of a depth map, $F_i(x, y)$ vector of elements in a fruxel model at slice i.

An object annotation is equal to a product of all elements with given x, y coordinates

$$A(x, y) = \prod_{i=0}^{d} F(x, y, i) \tag{3}$$

We use boolean operations to generate the fruxel model from a 3D scene. Firstly, we set the desired position of a virtual camera. After that, we find a boolean intersection between the 3D scene and XY slices of the frustum space. We render each intersection using white emission shader. We combine all slices in a single 3D array with dimensions $w \times h \times d$, where w is the width, h is the height of the color image, d is the number of slices. We term the resulting 3D array as a fruxel model. We generate fruxel models for real photos using 3D models generated with the structure-from-motion (SfM) algorithm. The SfM approach provides an estimation of camera poses with respect to the reconstructed 3D model. We place the virtual camera in the estimated pose and render the slices of the reconstructed model.

4.2 Conditional Adversarial Networks

Generative Adversarial Networks (GAN) generate a signal $\hat{B}$ for a given random noise vector z, $G : z \rightarrow \hat{B}$ [18,27]. Conditional GAN transforms an input image A and the vector z to an output $\hat{B}$, $G : \{A, z\} \rightarrow \hat{B}$. The input A can be an image that is transformed by the generator network G. The discriminator network D is trained to distinguish "real" signals from target domain B from the "fakes" $\hat{B}$ produced by the generator. Both networks are trained simultaneously. Discriminator provides the adversarial loss that enforces the generator to produce "fakes" $\hat{B}$ that cannot be distinguished from "real" signal B.

We train a generator $G : \{A\} \rightarrow \hat{B}$ to synthesize a fruxel model $\hat{B} \in \mathbb{R}^{w \times h \times d}$ conditioned by a color image $A \in \mathbb{R}^{w \times h \times 3}$.

4.3 Z-GAN Framework

We use `pix2pix` [27] framework as a starting point to develop our `Z-GAN` model. We keep the encoder part of the generator unchanged. We change 2D convolution kernels with 3D deconvolution kernels to encode a correlation between neighbor slices along the Z-axis.

We keep the skip connections between the layers of the same depth that was proposed in the U-net model [46]. We believe that skip connections help to transfer high-frequency components of the input image to the high-frequency components of the 3D shape. The resulting architecture of our `Z-GAN` model is presented in Fig. 6.

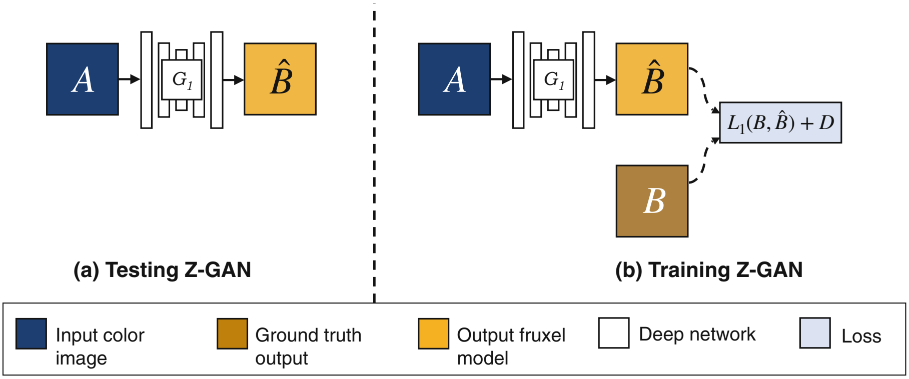


Fig. 6. `Z-GAN` framework.

4.4 Volumetric Generator

The main idea of our volumetric generator G is to use the correspondence between silhouettes in a color image and slices of a fruxel model. We used the U-Net generator [46] as a starting point to develop our model. The original U-Net generator leverages skip connections between convolutional and deconvolutional layers of the same depth to transfer fine details from the source to the target domain effectively.

We added two contributions to the original U-Net model. Firstly, we replaced the 2D deconvolutional filters with 3D deconvolutional filters. Secondly, we modified the skip connections to provide the correspondence between shapes of 2D and 3D features. The outputs of 2D convolutional filters in the left (encoder) side of Z-Net generator are $F_{2D} \in \mathbb{R}^{w \times h \times c}$ tensors, where w, h is the width and the height of a feature map and c is the number of channels. The output of 3D deconvolutional filters in the right (decoder) side are $F_{3D} \in \mathbb{R}^{w \times h \times d \times c}$ tensors. We use d copies of each channel of F_{2D} to fill the third dimension of F_{3D}. We term this operation as "copy inflate". The architecture of the generator is presented in Fig. 7.

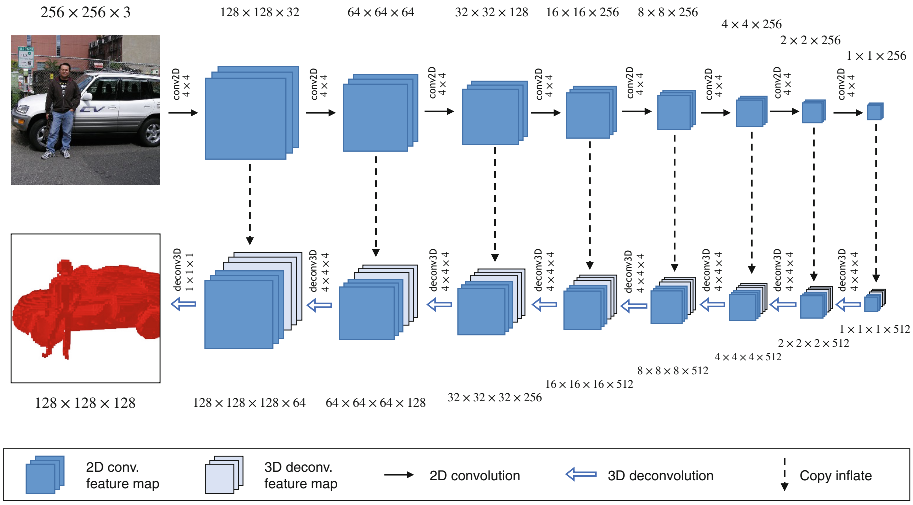


Fig. 7. The architecture of the generator.

4.5 Volumetric Discriminator

We modify the PatchGAN discriminator [27] to process the 3D slices efficiently. The original PatchGAN discriminator is based on the assumption of the markovian independence of the local image patches. Therefore the discriminator penalizes the image structure only at the scale of local patches.

The PatchGAN discriminator consists of a stack of convolutional layers with a constant kernel size. The stride of each layer is balanced with a kernel size in such way that the layer output size remains corresponding to the size of the input image. In other words, each convolutional layer takes the input with the size equal to the size of the input color image and produces a feature map. The sequential application of the convolutions with constant kernel size increases the "aperture" of the discriminator. For example, sequential application of seven convolutional layers results in the feature "aperture" of 140 pixels.

Our Z-Patch discriminator has a similar structure to the PatchGAN discriminator [27]. We replaced all 2D convolutional layers with 3D convolutional layers to process 3D shapes.

5 Evaluation

We evaluate baseline models and our `Z-GAN` framework on a task of generation of a voxel model from a single-view color image. We use two 3D shape datasets for the evaluation: Pascal 3D+ [60] and Pix3D [52]. All datasets include real images with 6D object poses.

We use two metrics to provide a quantitative evaluation of 3D object reconstruction quality: (*i*) an Intersection over Union (IoU) metric to measure a difference between a ground-truth 3D model and an output of a method and (*ii*) a surface distance metric similar to [45] to evaluate an accuracy of camera pose estimation for the `3D-R2N2` and our `Z-GAN` models. We also provide images of resulting voxel models for qualitative evaluation.

5.1 Baselines

We compare our model with three baselines: `3D-R2N2` [9,48], `TL-network` [17], and `MarrNet` [58]. To the best of our knowledge, there are no baselines to date capable of predicting voxel models of multiple objects from a single image. `TL-network` and `MarrNet` perform object-centered [48] prediction of voxel models with resolutions of $20 \times 20 \times 20$ and $128 \times 128 \times 128$. `3D-R2N2` provides a view-centered prediction with resolution $32 \times 32 \times 32$. Our `Z-GAN` model predicts a view-centered fruxel model with resolution $128 \times 128 \times 128$.

5.2 Training Details

Our `Z-GAN` framework was trained on the VoxelCity and VoxelHome datasets using PyTorch library [37]. We use VoxelCity dataset for the evaluation on Pascal 3D+ with fruxel model parameters $\{z_n = 2, z_f = 12, d = 128, \alpha = 40°\}$. For the evaluation on Pix3D dataset, we train our model on the VoxelHome dataset with fruxel model parameters $\{z_n = 0.5, z_f = 5.5, d = 128, \alpha = 40°\}$. The training was performed using the NVIDIA 1080 Ti GPU and took 11 h for G, D. For network optimization, we use minibatch SGD with an Adam solver. We set the learning rate to 0.0002 with momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$ similar to [27].

5.3 3D Reconstruction on Pascal 3D+

Qualitative Evaluation. We show results of single-view voxel model generation in Fig. 8. We use three object classes: car, bicycle, human. We selected 2,762 images from Pascal 3D+ image sets with a field of view similar to our trained model. We manually annotated the images with human 3D models from the ShapeNet dataset [8]. The qualitative evaluation demonstrates that models predicted by `TL-network` and `MarrNet` models have limited resolution and do not demonstrate new details compared to ground-truth models from the training set. While `3D-R2N2` shows more diversity in the output, it is capable of predicting only a single object in a scene. Our `Z-GAN` model produces voxel models of the whole scene with multiple object instances.

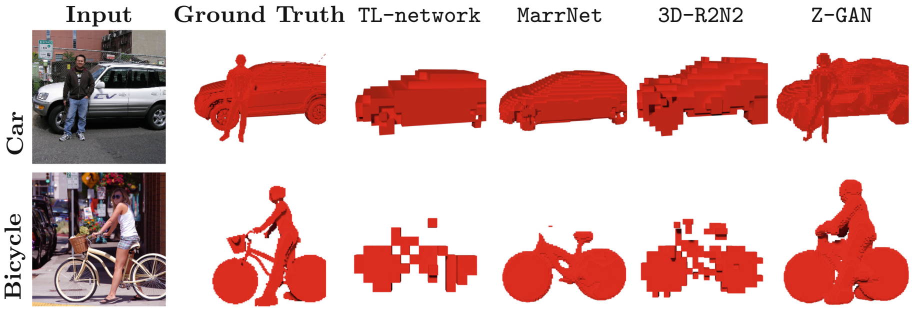


Fig. 8. An example of 3D reconstruction using `TL-network`,`MarrNet`,`3D-R2N2` and `Z-GAN` on Pascal 3D+ [60] dataset and considering three object classes: car, bicycle, human.

Quantitative Evaluation. We evaluate the results of the proposed `Z-GAN` method in terms of IoU and surface distance in Tables 3 and 4.

Table 3. Intersection over union metric for different object classes for Pascal 3D+ images.

Method	Object class		
	Car	Bicycle	Mean
`TL-network` [17]	0.301	0.117	0.209
`MarrNet` [58]	0.321	0.156	0.239
`3D-R2N2` [48]	0.582	0.212	0.397
`Z-GAN`	**0.612**	**0.398**	**0.505**

Table 4. Surface distance metric [45] for different object classes for Pascal 3D+ images.

Method	Object class		
	Car	Bicycle	Mean
`3D-R2N2` [48]	0.151	0.701	0.426
`Z-GAN`	**0.091**	**0.356**	**0.224**

5.4 3D Reconstruction on Pix3D

Qualitative Evaluation. Evaluation results of single-view voxel model generation are presented in Fig. 9. We use two object classes: chair and table. We selected 1,512 images from Pix3D image sets with a field of view similar to our model trained on VoxelHome dataset. We made the following conclusions from the qualitative evaluation. Firstly, `TL-network` predicts the object as the voxel model from the training set. While `MarrNet` tries to imitate the shape of the object in the input, it is confused on images with multiple objects. `3D-R2N2` reconstructs view-centered object voxel model but the resolution of the model is not enough to show details of multiple objects. Results of our `Z-GAN` model demonstrate fine object details and correct poses of multiple objects.

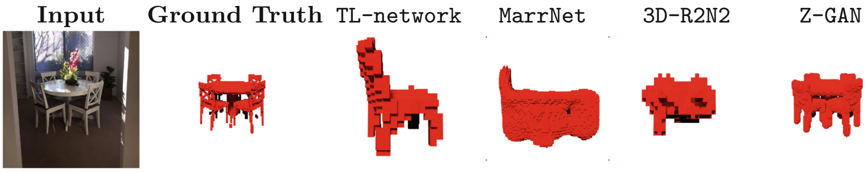


Fig. 9. Example of 3D reconstructions using the Pix3D dataset.

Quantitative Evaluation. We evaluate the results of the proposed `Z-GAN` method in terms of IoU and surface distance in Tables 5 and 6.

Table 5. Intersection over union metric for different object classes for Pix3D images.

Method	Object class		
	Chair	Table	Mean
`TL-network` [17]	0.190	0.211	0.201
`MarrNet` [58]	0.241	0.376	0.309
`3D-R2N2` [48]	0.289	0.251	0.270
`Z-GAN`	**0.461**	**0.612**	**0.536**

Table 6. Surface distance metric [45] for different object classes for Pix3D images.

Method	Object class		
	Chair	Table	Mean
`3D-R2N2` [48]	0.201	0.691	0.446
`Z-GAN`	**0.121**	**0.467**	**0.294**

6 Conclusions

The paper presented a new approach based on conditional generative adversarial networks capable of prediction of a voxel model from a single image. We showed that conditional adversarial volumetric networks can generate voxel models of complex scenes with multiple objects. We demonstrated that skip connections between 2D convolutional and 3D deconvolutional layers facilitate reconstruction of fine details. Furthermore, models utilizing skip connections require less training parameters for high-quality reconstruction of cluttered scenes with multiple 3D shapes of different classes.

We developed a new Z-GAN framework for translation of a single color image to a voxel model of a scene. We collected two datasets VoxelCity and VoxelHome to train our model. Datasets include fine-grade scene models, color images, depth maps and 6D object poses. We evaluated baselines and our model on multiple 3D shape datasets to show that it achieves and surpasses the state-of-the-art in terms of the number of reconstructed objects and their details.

Acknowledgments. The reported study was funded by Russian Foundation for Basic Research (RFBR) according to the research project N° 17-29-04509 and the Russian Science Foundation (RSF) according to the research project N° 16-11-00082.

We would like to thank the volunteers, who acted as statues for 3D models for class "human": Zoya Kniaz, Lena Metelkina, Nastya Metelkina, and Anya Metelkina. Also we would like to thank the authors of 3D CAD models for the dataset: Artyom Bordodymov and Pyotr Moshkantsev.

References

1. Alhaija, H.A., Mustikovela, S.K., Mescheder, L., Geiger, A., Rother, C.: Augmented reality meets computer vision: efficient data generation for urban driving scenes. Int. J. Comput. Vis. (2018). https://doi.org/10.1007/s11263-018-1070-x
2. Balntas, V., Doumanoglou, A., Sahin, C., Sock, J., Kouskouridas, R., Kim, T.: Pose guided RGBD feature learning for 3D object pose estimation. In: 2017 IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October, pp. 3876–3884 (2017). https://doi.org/10.1109/ICCV.2017.416
3. Balntas, V., Doumanoglou, A., Sahin, C., Sock, J., Kouskouridas, R., Kim, T.K.: Pose guided RGBD feature learning for 3D object pose estimation. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
4. Behl, A., Hosseini Jafari, O., Karthik Mustikovela, S., Abu Alhaija, H., Rother, C., Geiger, A.: Bounding boxes, segmentations and object coordinates: how important is recognition for 3D scene flow estimation in autonomous driving scenarios? In: The IEEE International Conference on Computer Vision (ICCV), October 2017
5. Brachmann, E., et al.: DSAC - differentiable RANSAC for camera localization. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
6. Brachmann, E., Rother, C.: Learning less is more - 6D camera localization via 3D surface regression. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
7. Brock, A., Lim, T., Ritchie, J., Weston, N.: Generative and discriminative voxel modeling with convolutional neural networks. pp. 1–9 December 2016. https://nips.cc/Conferences/2016. workshop contribution; Neural Inofrmation Processing Conference : 3D Deep Learning, NIPS; Conference date: 05–12-2016 Through 10–12-2016
8. Chang, A.X., et al.: ShapeNet: an information-rich 3D model repository. CoRR abs/1512.03012 (2015)

9. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3D–R2N2: a unified approach for single and multi-view 3D object reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
10. Doumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.: Recovering 6D object pose and predicting next-best-view in the crowd. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 3583–3592 (2016). https://doi.org/10.1109/CVPR.2016.390
11. Drost, B., Ulrich, M., Bergmann, P., Hartinger, P., Steger, C.: Introducing MVTec ITODD - a dataset for 3D object recognition in industry. In: The IEEE International Conference on Computer Vision (ICCV) Workshops, October 2017
12. El-Hakim, S.: A flexible approach to 3D reconstruction from single images. In: ACM SIGGRAPH, vol. 1, pp. 12–17 (2001)
13. Engel, J., Stueckler, J., Cremers, D.: Large-scale direct slam with stereo cameras (2015)
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2009)
15. Firman, M., Mac Aodha, O., Julier, S., Brostow, G.J.: Structured prediction of unobserved voxels from a single depth image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
16. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. (IJRR) **32**(11), 1231–1237 (2013)
17. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 484–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_29
18. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
19. Heinly, J., Schonberger, J.L., Dunn, E., Frahm, J.M.: Reconstructing the world* in six days *(as captured by the Yahoo 100 million image dataset). In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
20. Hinterstoisser, S., et al.: Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 548–562. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_42
21. Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: an RGB-D dataset for 6D pose estimation of texture-less objects. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2017)
22. Hodan, T., Haluza, P., Obdrzálek, S., Matas, J., Lourakis, M.I.A., Zabulis, X.: T-LESS: an RGB-D dataset for 6D pose estimation of texture-less objects. In: 2017 IEEE Winter Conference on Applications of Computer Vision WACV 2017, Santa Rosa, CA, USA, 24–31 March 2017, pp. 880–888 (2017). https://doi.org/10.1109/WACV.2017.103
23. Hodaň, T., Matas, J., Obdržálek, Š.: On evaluation of 6D object pose estimation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 606–619. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_52
24. Hodaň, T., Michel, F., Sahin, C., Kim, T.K., Matas, J., Rother, C.: SIXD Challenge 2017. http://cmp.felk.cvut.cz/sixd/challenge_2017/. Accessed 01 July 2018

25. Hoppe, C., Klopschitz, M., Donoser, M., Bischof, H.: Incremental surface extraction from sparse structure-from-motion point clouds. In: Proceedings of the British Machine Vision Conference 2013, pp. 94:1–94:11, British Machine Vision Association (2013)
26. Huang, Q., Wang, H., Koltun, V.: Single-view reconstruction via joint analysis of image and shape collections. ACM Trans. Graph. **34**(4), 87:1–87:10 (2015)
27. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976. IEEE (2017)
28. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In: Proceedings of the IEEE International Conference on Computer Vision, University of Cambridge, Cambridge, United Kingdom, pp. 2938–2946. IEEE, February 2015
29. Kniaz, V.V.: Robust vision-based pose estimation algorithm for an UAV with known gravity vector. ISPRS-Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci. **XLI-B5**, 63–68 (2016). https://doi.org/10.5194/isprs-archives-XLI-B5-63-2016
30. Knyaz, V., Zheltov, S.: Accuracy evaluation of structure from motion surface 3D reconstruction. In: Proceedings of SPIE, vol. 10332, pp. 10332-1–10332-10 (2017). https://doi.org/10.1117/12.2272021
31. Knyaz, V.A., et al.: Deep learning of convolutional auto-encoder for image matching and 3d object reconstruction in the infrared range. In: The IEEE International Conference on Computer Vision (ICCV) Workshops, pp. 2155–2164 (2017). https://doi.org/10.1109/ICCVW.2017.252
32. Krull, A., Brachmann, E., Nowozin, S., Michel, F., Shotton, J., Rother, C.: PoseAgent: budget-constrained 6d object pose estimation via reinforcement learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
33. Lim, J.J., Pirsiavash, H., Torralba, A.: Parsing IKEA objects: fine pose estimation. In: ICCV (2013)
34. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440. IEEE (2015)
35. Ma, M., Marturi, N., Li, Y., Leonardis, A., Stolkin, R.: Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos. Pattern Recogn. **76**(11), 506–521 (2017)
36. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR, pp. 3061–3070 (2015)
37. Paszke, A., et al.: Automatic differentiation in pyTorch (2017)
38. Poiesi, F., Locher, A., Chippendale, P., Nocerino, E., Remondino, F., Van Gool, L.: Cloud-based collaborative 3D reconstruction using smartphones. In: the 14th ACM European Conference on Visual Media Production (CVMP), pp. 1–9. ACM Press, New York (2017)
39. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointNets for 3D object detection from RGB-D data. arXiv preprint arXiv:1711.08488 (2017)
40. Rad, M., Lepetit, V.: BB8: a scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 3848–3856 (2017). https://doi.org/10.1109/ICCV.2017.413

41. Remondino, F., Nocerino, E., Toschi, I., Menna, F.: A critical review of automated photogrammetric processing of large datasets. ISPRS - Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci. **42**, 591–599 (2017). XLII-2/W5. https://doi.org/10.5194/isprs-archives-XLII-2-W5-591-2017
42. Remondino, F., Roditakis, A.: Human figure reconstruction and modeling from single image or monocular video sequence. In: 2003 Fourth International Conference on 3-D Digital Imaging and Modeling, 3DIM 2003, pp. 116–123. IEEE October 2003
43. Remondino, F., El-Hakim, S.: Image-based 3D modelling: a review. Photogram. Rec. **21**(115), 269–291 (2006)
44. Richter, S.R., Roth, S.: Matryoshka networks: predicting 3D geometry via nested shape layers. arXiv.org, April 2018
45. Rock, J., Gupta, T., Thorsen, J., Gwak, J., Shin, D., Hoiem, D.: Completing 3D object shape from one depth image. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2484–2493. University of Illinois at Urbana-Champaign, Urbana, IEEE, October 2015
46. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
47. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016
48. Shin, D., Fowlkes, C., Hoiem, D.: Pixels, voxels, and views: a study of shape representations for single view 3D object shape prediction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
49. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in RGB-D images. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pp. 2930–2937. IEEE Computer Society, Washington (2013). https://doi.org/10.1109/CVPR.2013.377
50. Sock, J., Kim, K.I., Sahin, C., Kim, T.K.: Multi-task deep networks for depth-based 6D object pose and joint registration in crowd scenarios. arXiv.org, June 2018
51. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
52. Sun, X., et al.: Pix3D: dataset and methods for single-image 3D shape modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
53. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Multi-view 3D models from single images with a convolutional network. arXiv.org, November 2015
54. Tefera, Y., Poiesi, F., Morabito, D., Remondino, F., Nocerino, E., Chippendale, P.: 3DNOW: image-based 3D reconstruction and modeling via web. ISPRS - Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci. 1097–1103 (2018). XLII-2. https://doi.org/10.5194/isprs-archives-XLII-2-1097-2018
55. Tejani, A., Kouskouridas, R., Doumanoglou, A., Tang, D., Kim, T.: Latent-class hough forests for 6 DoF object pose estimation. IEEE Trans. Pattern Anal. Mach. Intell. **40**(1), 119–132 (2018). https://doi.org/10.1109/TPAMI.2017.2665623
56. Valentin, J., et al.: Learning to navigate the energy landscape. In: Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016, University of Oxford, Oxford, United Kingdom, pp. 323–332. IEEE, December 2016

57. Walas, K., Nowicki, M., Ferstl, D., Skrzypczynski, P.: Depth data fusion for simultaneous localization and mapping - RGB-DD SLAM. In: 2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, MFI 2016, Baden-Baden, Germany, 19–21 September 2016, pp. 9–14 (2016). https://doi.org/10.1109/MFI.2016.7849459
58. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W.T., Tenenbaum, J.B.: MarrNet: 3D shape reconstruction via 2.5D sketches. arXiv.org November 2017
59. Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling, pp. 82–90 (2016)
60. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond PASCAL: a benchmark for 3D object detection in the wild. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2014)
61. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3D object reconstruction without 3D supervision (2016). papers.nips.cc
62. Yang, B., Rosa, S., Markham, A., Trigoni, N., Wen, H.: 3D object dense reconstruction from a single depth view. arXiv preprint arXiv:1802.00411 (2018)
63. Yang, B., Wen, H., Wang, S., Clark, R., Markham, A., Trigoni, N.: 3D object reconstruction from a single depth view with adversarial learning. In: The IEEE International Conference on Computer Vision (ICCV) Workshops, October 2017
64. Zheng, B., Zhao, Y., Yu, J.C., Ikeuchi, K., Zhu, S.C.: Beyond point clouds: scene understanding by reasoning geometry and physics. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2013