RGBD BASED GENERATIVE ADVERSARIAL NETWORK FOR 3D SEMANTIC SCENE COMPLETION

Jiahao Wang, Ling Pei*, Danping Zou, Yifan Zhu,
Tao Li and Ruochen Wang

Shanghai Key Laboratory of Location-based Navigation and Services
SJTU-ParisTech Elite Institute of Technology
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

3D scene understanding is of importance since it is a reflection about the real-world scenario. The goal of our work is to complete the 3d semantic scene from an RGB-D image. The state-of-the-art methods have poor accuracy in the face of complex scenes. In addition, other existing 3D reconstruction methods use depth as the sole input, which causes performance bottlenecks. We introduce a two-stream approach that uses RGB and depth as input channels to a novel GAN architecture to solve this problem. Our method demonstrates excellent performance on both synthetic SUNCG and real NYU dataset. Compared with the latest method SSCNet, we achieve 4.3% gains in Scene Completion (SC) and 2.5% gains in Semantic Scene Completion (SSC) on NYU dataset.

KEYWORDS

Scene Completion, Semantic Segmentation, Generation Adversarial Network, RGB-D

1. INTRODUCTION

We live in a three-dimensional world. In order to live in the three-dimensional world and interact with the external environment, humans rely on the observation and analysis of the 3D geometry and semantics of surrounding objects. Similarly, the ability to infer complete 3D shapes from local observations is essential for robots, it can achieve low-level tasks such as grasping and avoiding obstacles. Moreover, the ability to infer the semantics of objects in a scene can achieve higher-level goals, such as the task of object retrieval. Therefore, in order to better sense the surrounding environment, the robot needs to construct a semantic scene map.

Semantic scene completion is a combined task of semantic segmentation and shape completion and discovers the hidden information existing in the 3D scene. For instance, a depth sensor can only capture information from object surfaces that are visible. Most of the geometric and semantic information of the 3D scene is, however, occluded by the objects themselves. As humans, we can estimate the geometry of objects even in the occluded area from experience, providing us instantly an effective model of the 3D scene surrounding us. 3D semantic scene completion tries to achieve the same goal. Given a single depth image, the goal is to predict the entire 3D geometry of all objects in the scene including the occluded areas. The technique has high potential in many areas ranging from domestic robotics and autonomous vehicles to health-care systems. However, due to the dimensional curse brought by 3D representation and the limited annotation datasets, the research field of semantic scene completion still step slowly in the past decades. To push the scientific effort along these directions, recently large-scale

Natarajan Meghanathan et al. (Eds) : CoSIT, AIAPP, SIGL, CRIS, DMA, NLPML, CYBI - 2020
pp. 117-125, 2020. © CS & IT-CSCP 2020 DOI: 10.5121/csit.2020.100111

benchmark datasets, such as SUNCG[1], NYU[2], ScanNet[3] and SceneNet,[4] have been proposed to evaluate different visual scene understanding tasks including those of scene completion and semantic segmentation.

Inspired by the way humans can imagine the 3D structure of a room by looking at a 2D image, we propose an algorithm that reconstructs the entire scene geometry and semantics from an RGB-D image. By directly reconstructing the scene from one view, the challenge is to plausibly complete the scene in place of the hidden structures that are not visible from the input RGB-D image. To this end, we utilize a learning strategy that allows the algorithm to simultaneously perceive the objects in the scene and use its contextual shape to fill the hidden structures. In addition, we simultaneously estimate a semantic segmentation of the completed 3D scene geometry.

In this work, we focus on the data acquired from RGB-D cameras, with the goal of reconstructing and semantically labelling the whole scene from one single range image. As a scene may contain small objects and complicated shapes, we apply a generative adversarial model for this semantic completion task. Combined with an encoder and a generator, our architecture uses both RGB images and depth images as the input information and generate 3D volumetric data whose elements are labelled with object categories. Specifically, we use two discriminators to train the architecture to back-project the colour and depth information into the 3D volumetric space with semantic labels. One discriminator is used to optimize the entire architecture by comparing the reconstructed semantic scene with the ground truth. Our generative adversarial network formulates the 3D scene completion and labelling as a joint task and learns in an end-to-end way. The main contributions of this paper are three-fold:

(1) We propose a novel generative adversarial network (GAN) to predict semantic labels and occupancy in 3D space simultaneously.
(2) 3D feature maps of RGB and depth are fused in multi-scale seamlessly, which enhances the network representation ability and boost the performance of SC and SSC tasks.
(3) The proposed end-to-end training network achieves state-of-the-art performance on SUNCG and NYU datasets.

2. RELATED WORKS

2.1. 3D Scene Analysis

A set of methods have been proposed for scene segmentation, scene completion, and object detection from an input RGBD image or depth image. 2D image-based methods regard the depth as an additional channel of the 2D RGB image and leverage manually-crafted features[5,6] or 2D deep neural networks for these scene analysis tasks[7,8]. 3D volume-based approaches convert the input depth map into a volumetric representation and exploit manually crafted 3D features[9] or 3D CNNs for detecting 3D objects from the input RGBD image[10]. Although these methods can successfully detect and segment visible 3D objects and scenes in the input RGBD images, they cannot infer the scenes that are totally occluded. Instead, our method predicts semantic labelling and 3D shapes for both visible and invisible objects in a 3D scene.

Liu et al.[11] introduced 3DCNN-DQN-RNN for parsing 3D point cloud of a scene. PointNet[12] and PointNet++[13] develop deep learning framework on 3D point cloud for scene semantic labelling and other 3D shape analysis tasks. These methods take the 3D point cloud of whole 3D scene as the input. On the contrary, our method takes a single depth image for semantic scene completion.

2.2. 3D Scene Completion

Firman et al.[14] inferred the occluded 3D object shapes from a single depth image via random forest. Zheng et al.[15] completed the occluded scene in the input depth image with a set of pre-defined rules and refined the completion results by physical reasoning. These methods perform scene segmentation and completion in two separate steps. Recently, Song et al.[16] proposed 3D SSCNet for simultaneously predicting the semantic labels and volumetric occupancy of the 3D objects from a single depth image. Although this method unifies segmentation and completion and significantly improves the result, the expensive 3D CNN limits the input volume resolution and network depth, and thus restrains its performance. By combining 2D CNN and 3D CNN, our method efficiently reduces the training and inference cost, enhances the network depth and thus significantly improves the result accuracy.

2.3. 3D Object Completion

A set of methods reconstruct the 3D object shape from a single depth image using 3D shape retrieval[17], Convolutional Deep Belief Network (CDBN)[18,19], or a 3D Generative Adversarial Networks (GAN)[20]. All these methods model the input depth maps and resulting 3D shapes with a 3D volumetric representation. Although these methods can be combined with other scene segmentation methods for predicting 3D shapes of the visible object in the input depth map, they cannot be used for inferring objects that are totally occluded. Our method is designed for recovering complete 3D shapes of both visible and occluded 3D objects from a single depth image of a 3D scene.

2.4. 3D Scene Completion and Semantic Labelling

3D semantic scene completion has become a popular research problem recently. Some prior works considered completing and labelling 3D scenes as a combined task, but they used separate modules for feature extraction and context modelling [20,21,22]. Song et al.[23] pioneered in applying deep learning to semantic scene completion. They proposed a 3D convolutional network that leverages dilated convolutions [24] as well as skip connections [25]. Also, this work has been extended by adding a second input stream which contains the 2D semantic labels from RGB images [26,27], the authors proposed a coarse-to-fine 3D fully convolutional network for processing 3D scenes with arbitrary spatial extents and capturing both local details and the global structure of the scenes [28].

3. RGBD BASED SEMANTIC SCENE COMPLETION WITH GANS

Inspired by the successful application of GANs in other domains, we introduce a novel model to perform semantic scene completion using GANs.

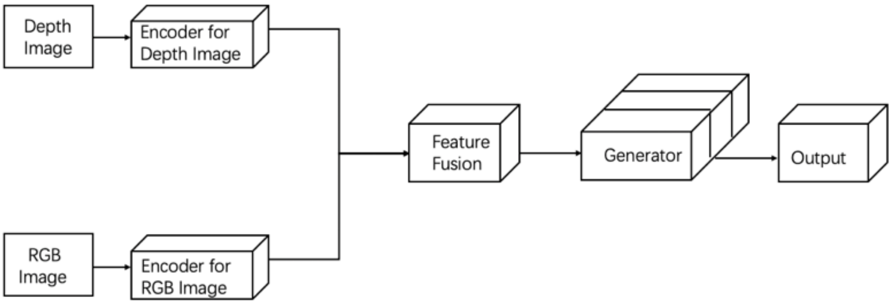


Figure 1. Proposed network architecture.

3.1. Network Architecture

From the RGB-D image to the 3D volume, our architecture is a concatenation of an encoder E_{dep} with 2D convolutional operators that convert the input depth image into a lower dimensional latent feature l_{dep}; and, an encoder E_{rgb} with 2D convolutional operators that convert the input colour image into a lower dimensional latent feature l_{rgb} ; a feature fusion module F_{fusion} that convert the latent feature l_{dep} and l_{rgb} into a fusion feature l_{fusion} ;a generator G with 3D deconvolutional kernels that takes l_{fusion} to build the semantic reconstruction. This architecture is illustrated in Figure. 1.

3.2. Encoder for depth image

The encoder E_{dep} compresses the depth image into a feature in the latent space. Its architecture is a concatenated network that sequentially combines 2D convolutional layers and max-pooling layers. The operators for the paired convolutional and pooling layers are 2D convolutional kernels with, respectively, the size of 3 * 3 and stride of 1 * 1 and the size of 2 * 2 with stride of 2 * 2. Each of these paired layers is processed by a leaky ReLU activation function. Therefore, the output of every ReLU activation is a multi-channel 2D image. After six convolutions operations, the result is an 80-channel 5 * 3 image. The output of the encoder represents the latent feature l_{dep} of the semantic reconstruction architecture.

3.3. Encoder for colour image

The encoder E_{rgb} is first processed by a 2D-CNN for semantic segmentation. The network is an adaptation of the Resnet101 architecture for semantic segmentation. While all but one pooling layer are omitted, dilated convolutions are used to keep the output resolution high while simultaneously increasing the receptive field. The output is down sampled by a factor of 4 with respect to the input. The output is then upsampled using bilinear interpolation. The 2D-CNN predicts the softmax probabilities for every class and pixel. The output of the encoder represents the latent feature l_{rgb} of the semantic reconstruction architecture.

3.4. Feature fusion module

We propose a novel feature fusion strategy which can fully use the multi-modal features. We employ multi-modal CNN feature fusion while preserving the lower computational cost. In specific, different levels of features are extracted through E_{rgb} and E_{dep}, and then these features are merged together by element-wise add. The reason for using element-wise add rather than other operations is because it can fuse the features neatly with insignificant computation costs.

3.5. Generator

With the goal of regressing the semantic reconstruction, the generator G unwraps the latent feature to a higher dimensional voxel data. We assemble the generator with 3D deconvolutional layers with the size of 3 * 3 * 3 and stride of 2 * 2 * 2 which are processed by the ReLU function as activation. After four deconvolutional layers, the output of the generator is the voxel-wise classification y. By doing this, y is presented in the shape of 80 * 48 * 80 * K, with K object classes.

3.6. Discriminator

The discriminator network consists of several convolutional blocks. Each block comprises a convolution layer with a 3D kernel, a normalization layer, and a leaky ReLU activation layer. The output of the last convolutional layer with the size of 5 * 3 * 5 * 16 is reshaped to a vector of 1200 dimensions. After that, it is processed by three fully-connected layers with output sizes of 256, 128 and 1, respectively. Hence, the final logit is a binary indicator to determine whether the predicted volumetric data is generated or sampled from the ground truth data.

4. Optimization

In this method we propose to use a hybrid loss function that is a weighted sum of two terms. The first term is a multi-class cross-entropy loss that is used for the generator to predict the right class label at each voxel location independently. We use g(x) to denote the class probability map over the C classes for the volume H * W * D, which is produced by the generator network. The second loss term is based on the output of the discriminator network. This loss term is large if the discriminator can differentiate between the predictions of the generator network and the ground truth label maps. We use d(x, y) ∈ [0, 1] to represent the probability with which the discriminator network predicts that y is the ground truth label map of x, as opposed to being a label map produced by the generator network g(·). Given a dataset of N training images x_n and a corresponding 3D ground truth volume y_n, we define the loss as:

$$\mathcal{L}_{GAN}\left(\theta_g, \theta_d\right) = \sum_{n=1}^{N} \mathcal{L}_{mce}\left(g\left(x_n\right), y_n\right) \quad -\lambda\left[\mathcal{L}_{bce}\left(d\left(x_n, y_n\right), 1\right) + \mathcal{L}_{bce}\left(d\left(x_n, g\left(x_n\right)\right), 0\right)\right] \tag{1}$$

where θ_g and θ_d denote the parameters of the generator and discriminator network, respectively. The multi-class cross-entropy loss for prediction y is given by:

$$\mathcal{L}_{mce}(\hat{y}, y) = -\sum_{i=1}^{H\times W\times D} \sum_{n=1}^{N} y_{in} \ln \hat{y}_{in} \tag{2}$$

which equals the negative log-likelihood of the target ground truth volume y in a one-hot encoding representation.

Similarly, the binary cross-entropy loss is denoted as:

$$\mathcal{L}_{bce}(\hat{z}, z) = -[z \ln \hat{z} + (1 - z)\ln(1 - \hat{z})] \tag{3}$$

We then minimize the loss according to the parameters θ_g of the generator network, while maximizing it with respect to the parameters θ_d of the discriminator network.

5. Experiments

We implement our network architecture in PyTorch and use a batch size of 4. For our generator network, we use a SGD optimizer with weight decay of 0.0005 and learning rate of 0.01. For the discriminator network, we use an Adam optimizer with a learning rate of 0.0001.

We separate our evaluation results mainly in two parts: Semantic scene completion (SSC) and scene completion (SC). While scene completion only considers whether a voxel is occupied or empty, semantic scene completion also evaluates whether an occupied voxel is given the correct semantic label. We measure the precision, recall, and Jaccard index (IoU) for scene completion and the average (avg.) of the IoU across all categories for semantic scene completion.

5.1. Evaluation on SUNCG

SUNCG is a dataset of 3D scenes which contains pairs of depth image and its corresponding volumetric scene where all objects in the scene are semantically annotated. We implemented the 10-fold validation on the pairs for the 111,697 different scenes.

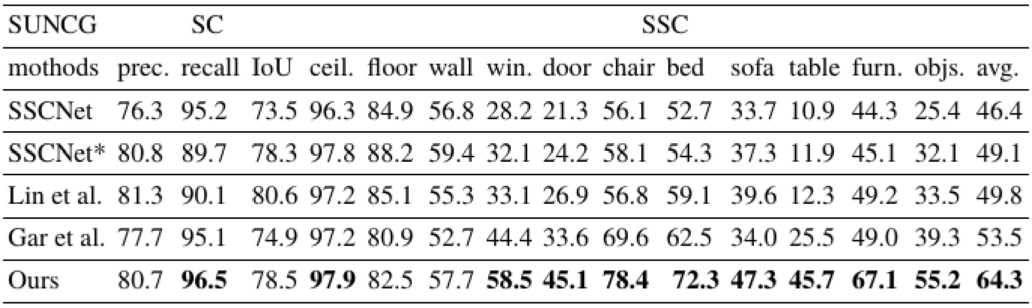
Table 1 shows the quantitative results of the SUNCG data set. We mainly compare our framework with the benchmark method (SSCNet) and its derivative methods. It can be seen that the 3D scene constructed by the SSCGAN network has a higher IoU, which is 78.5 \%; for the semantic scene reconstruction task, SSCGAN shows a better generation effect on the objects in the scene, especially for chairs, Smaller objects such as beds, sofas and furniture. The final average value (avg.) of IoU increased from 46.4 \% to 64.3 \%.

5.2. Evaluation on NYU

NYU dataset which is also an indoor scene dataset. It contains both the depth images captured by Kinect and the 3D models. This includes the volumetric 3D data with the annotated object labels for every voxels in 1,449 scenes. The semantic annotations for the volumetric data in this dataset consist of 33 objects in 7 categories.

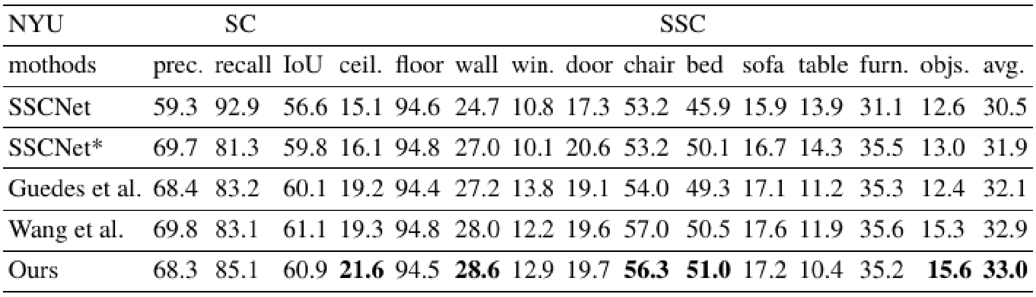
Table 2 shows the quantitative results of the NYU dataset. Although for 3D scene reconstruction tasks, the improvement brought by the SSCGAN network is not very obvious. However, for semantic segmentation tasks, the introduction of generative adversarial networks and RGB image streams enhances understanding of the details in the scene. The average IoU (avg.) Increased from 30.5 \% to 33.0 \%.

Table 1. Results on the SUNCG dataset.



SUNCG	SC			SSC										
mothods	prec.	recall	IoU	ceil.	floor	wall	win.	door	chair	bed	sofa	table	furn.	objs.	avg.
SSCNet	76.3	95.2	73.5	96.3	84.9	56.8	28.2	21.3	56.1	52.7	33.7	10.9	44.3	25.4	46.4
SSCNet*	80.8	89.7	78.3	97.8	88.2	59.4	32.1	24.2	58.1	54.3	37.3	11.9	45.1	32.1	49.1
Lin et al.	81.3	90.1	80.6	97.2	85.1	55.3	33.1	26.9	56.8	59.1	39.6	12.3	49.2	33.5	49.8
Gar et al.	77.7	95.1	74.9	97.2	80.9	52.7	44.4	33.6	69.6	62.5	34.0	25.5	49.0	39.3	53.5
Ours	80.7	**96.5**	78.5	**97.9**	82.5	57.7	**58.5**	**45.1**	**78.4**	**72.3**	**47.3**	**45.7**	**67.1**	**55.2**	**64.3**

Table 2. Results on the NYU dataset.



NYU	SC			SSC											
mothods	prec.	recall	IoU	ceil.	floor	wall	win.	door	chair	bed	sofa	table	furn.	objs.	avg.
SSCNet	59.3	92.9	56.6	15.1	94.6	24.7	10.8	17.3	53.2	45.9	15.9	13.9	31.1	12.6	30.5
SSCNet*	69.7	81.3	59.8	16.1	94.8	27.0	10.1	20.6	53.2	50.1	16.7	14.3	35.5	13.0	31.9
Guedes et al.	68.4	83.2	60.1	19.2	94.4	27.2	13.8	19.1	54.0	49.3	17.1	11.2	35.3	12.4	32.1
Wang et al.	69.8	83.1	61.1	19.3	94.8	28.0	12.2	19.6	57.0	50.5	17.6	11.9	35.6	15.3	32.9
Ours	68.3	85.1	60.9	**21.6**	94.5	**28.6**	12.9	19.7	**56.3**	**51.0**	17.2	10.4	35.2	**15.6**	**33.0**

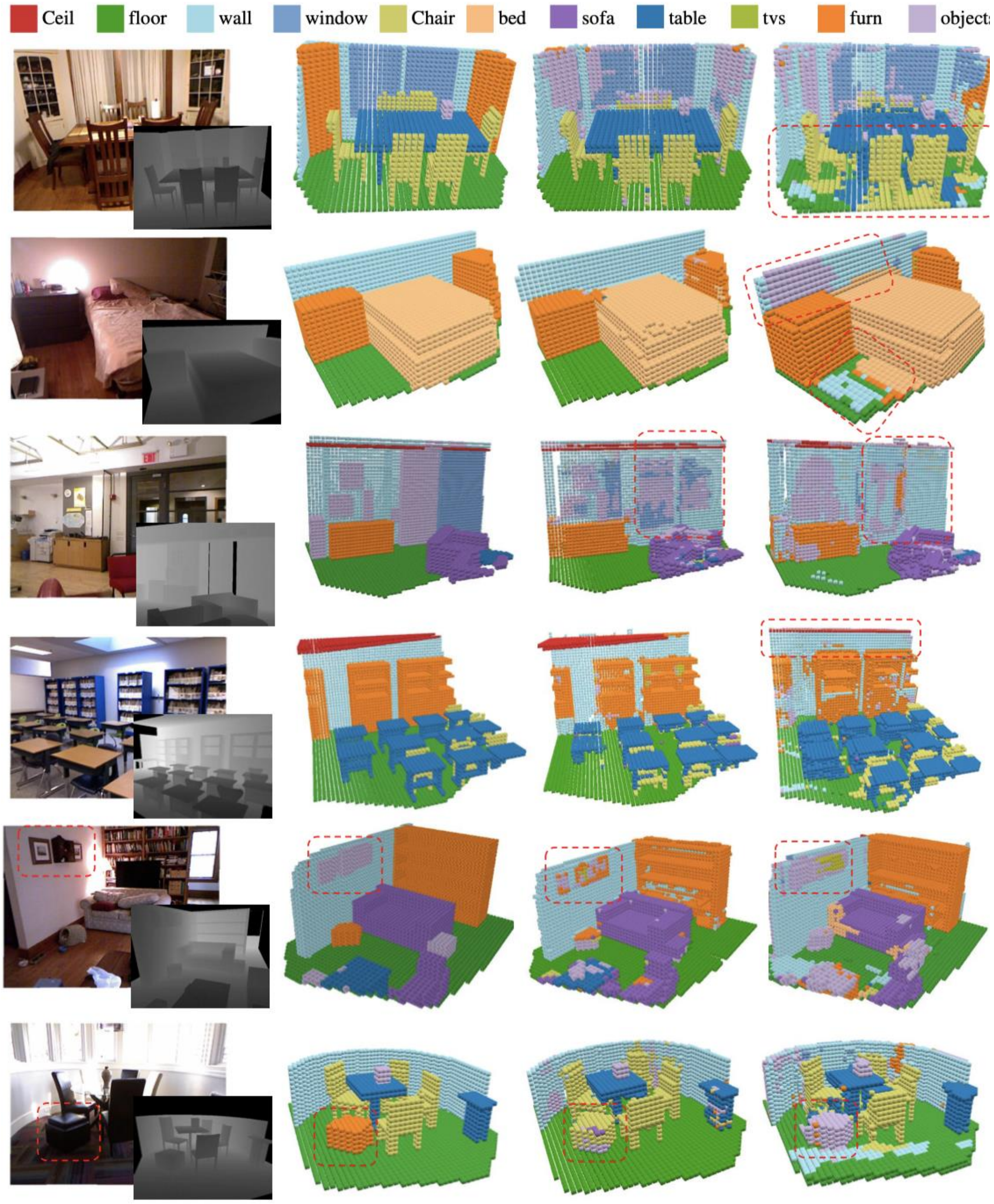


Figure 2. Qualitative Results

Figure 2. **Qualitative Results**. From left to right: Input RGB-D image, ground truth, results obtained by our approach, and results obtained by SSCNet . The scene completion results of our method are richer in details (marked with red dashed lines in the figure), and less prone to errors.

6. CONCLUSION

We presented a novel GAN architecture to perform 3D semantic scene completion base on an RGB-D image. The results show that the RGBD-GAN improves the network performance on both test sets. In comparison to the baseline, our models yield a significant improvement on

NYU v2 dataset. On SUNCG our models outperform the baseline by a large margin. If we compare the results qualitatively, the proposed model produces significantly more realistic appearing scenes than the baseline.

ACKNOWLEDGEMENTS

This work was supported by the National Nature Science Foundation of China (NSFC) under Grant 61873163, Equipment Pre-Research Field Foundation under Grant 61405180205, Grant 61405180104.

REFERENCES

[1] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[2] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In European Conference on Computer Vision, pages 746–760. Springer, 2012.

[3] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), volume 1, 2017.

[4] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla. Scenenet: Understanding real world indoor scenes with synthetic data. arXiv preprint arXiv:1511.07041, 2015.

[5] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. In CVPR. 2013.

[6] A. Atapour-Abarghouei and T.P. Breckon. Depthcomp: Real-time depth image completion based on prior semantic scene segmentation. In BMVC, pages 1–13, 2017.

[7] Saurabh Gupta, Ross Girshick, Pablo Arbel´aez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In ECCV, pages 345–360, 2014.

[8] Saurabh Gupta, Pablo Andr´es Arbel´aez, Ross B. Girshick, and Jitendra Malik. Aligning 3D models to RGB-D images of cluttered scenes. In CVPR, pages 4731–4740, 2015.

[9] Zhile Ren and Erik B. Sudderth. Three-dimensional object detection and layout prediction using clouds of oriented gradients. In CVPR, pages 1525–1533, 2016.

[10] Shuran Song and Jianxiong Xiao. Deep Sliding Shapes for amodal 3D object detection in RGB-D images. In CVPR, pages 808–816, 2016.

[11] Fangyu Liu, Shuaipeng Li, Liqiang Zhang, Chenghu Zhou, Rongtian Ye, Yuebin Wang, and Jiwen Lu. 3dcnn-dqn-rnn: A deep reinforcement learning framework for semantic parsing of large-scale 3d point clouds. In ICCV, pages 5678–5687, 2017.

[12] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In CVPR, pages 652–660, 2016.

[13] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In NIPS, pages 5105–5114, 2017.

[14] Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J. Brostow. Structured prediction of unobserved voxels from a single depth image. In CVPR, pages 54315440, 2016.

[15] Bo Zheng, Yibiao Zhao, Joey C. Yu, Katsushi Ikeuchi, and Song-Chun Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In CVPR, pages 3127–3134, 2013.

[16] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In CVPR, pages 17461754, 2017.

[17] Jason Rock, Tanmay Gupta, Justin Thorsen, JunYoung Gwak, Daeyun Shin, and Derek Hoiem. Completing 3d object shape from one depth image. In CVPR, pages 24842493, 2015.

[18] Zhirong Wu, Shuran Song, Aditya Khosla, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shape modeling. In CVPR, pages 1912–1920, 2015.

[19] D. T. Nguyen, B. S. Hua, M. K. Tran, Q. H. Pham, and S. K. Yeung. A field model for repairing 3d shapes. In CVPR, pages 5676–5684, 2016.

[20] Bo Yang, Hongkai Wen, Sen Wang, Ronald Clark, Andrew Markham, and Niki Trignoi. 3d object reconstruction from a single depth view with adversarial learning. In ICCV, pages 679–688, 2017.

[21] Bo Zheng, Yibiao Zhao, Joey C Yu, Katsushi Ikeuchi, and Song-Chun Zhu, “Beyond point clouds: Scene understanding by reasoning geometry and physics,” in CVPR, 2013, pp. 3127–3134.

[22] Byung-soo Kim, Pushmeet Kohli, and Silvio Savarese, “3D scene understanding by voxel-CRF,” in ICCV, 2013, pp. 14251432.

[23] Maros Blaha, Christoph Vogel, Audrey Richard, Jan D Wegner, Thomas Pock, and Konrad Schindler, “Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multiclass volumetric labeling,” in CVPR, 2016, pp. 3176–3184.

[24] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser, “Semantic scene completion from a single depth image,” in CVPR, 2017, pp. 190–198.

[25] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” TPAMI, vol. 40, no. 4, pp. 834–848, 2018.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.

[27] Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall, “Two Stream 3D Semantic Scene Completion,” in CVPR Workshops, 2019.

[28] Shice Liu, YU HU, Yiming Zeng, Qiankun Tang, Beibei Jin, Yainhe Han, and Xiaowei Li, “See and Think: Disentangling Semantic Scene Completion,” in NIPS, 2018, pp. 261–272.