Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

AM^{2}FNet: Attention-based Multiscale & Multi-modality Fused Network

Rong Chen
AI Institute
Ecovacs Robotics
Nanjing, Jiangsu Province, China
37110060@qq.com

Zhiyong Huang and Yuanlong Yu*
AI Institute
Ecovacs Robotics
Nanjing, Jiangsu Province, China
{*nn.huang, jeff.yu*}*@ecovacs.com*

***Abstract*—How to infer the 3D geometries and 3D semantic labels for each unit in a scene, including visible surfaces and occluded parts, is an important issue in many robotic fields. In recent years, there exists some studies on segmenting and completing 3D scene from 2D information. Most of them complete a scene from a single depth image. Compared with the depth image, the RGB image contains more color features and contour features, which can help to semantic labeling. However, how to design an effective strategy to fuse RGB and depth features is a challenge issue. Our paper presents an attention-based multi-scale & multi-modality fused network, called AM^{2}FNet, which includes six modules: depth feature module, color feature module, 3D integration module for multi-modality feature fusion, 3D refinement module for multi-scale feature fusion, attention modules, semantic mapping module. The integration module and the refinement module work together in 3D space to fuse color and depth features at low-level, middle-level and high-level in a top-down fashion. In addition, we use an attention module to efficiently bias input-related features. Experimental results show that our proposed network can generate higher-quality semantic scene completion (SSC) results and scene completion (SC) results, and outperforms the state-of-the-art methods on real NYU and synthetic NYUCAD datasets. Meanwhilethe contribions of single modules have been illustrated.**

***Index Terms*—3D integration module, 3D refinement module, attention, top-down fashion**

I. Introduction

3D semantic scene completion (SSC) is one of important functions for robots. It aims to complete the 3D scene representation and recognize their semantic labels for all voxels of the 3D scene. Based on such function, robots can accomplish more complex tasks.

Recently, with the development of deep learning techniques and large-scale datasets, there have been achievements in 3D SSC. One type of such existing 3D SSC methods is based on depth image only [1]–[3]. Song et al. [2] proposed a method called SSCNet, which originally presented a 3D convolutional neural network (CNN) to complete and label 3D scene simultaneously by integrating multiscale feature concatenations, shortcut connections and dilated convolutions. Thus SSCNet can achieve high performance in terms of scene completion. However, there are still several issues for such methods. First, these methods cannot retain local details such that the obtained semantic segmentations are discontinuous and contour is coarse. Secondly, semantic predictions of SSCNet are wrong in some cases. For example, SSCNet gives the same semantic labels for the regions with the same depth whereas different colors, as shown in Fig.1. Therefore, it is reasonable to integrate the RGB color information into 3D SSC in order to facilitate semantic labeling. The third issue is that these methods might output input-irrelative semantic label, e.g., a region might be labeled as a chair(yellow) but there is no chair in the input scene, as shown in the third row of Fig.1.

However, how to combine color and depth information is challenging. Some methods [4]–[6] extract color feature representation and predict semantic segmentation in 2D space, both of which are then integrated into the depth feature obtained in 3D space to construct an overall semantic feature representation. During the integration process, the color semantic feature is projected into 3D space and then concatenated with or added to depth feature. However, based on the fact that inferring 3D geometries and semantic labels are highly correlated, the first issue of such methods is that 2D semantic segmentation lacks 3D geometry information. The second issue is that color and depth information cannot be fused at multiple scales and stages since the integration is executed after 2D segmentation rather than at various stages in such methods.

In order to cope with aforementioned issues, this paper proposes an attention-based multiscale & multi-modality fused network (AM^{2}FNet), which includes six modules: depth feature module, color feature module, 3D integration module for multi-modality feature fusion, 3D refinement module for multi-scale feature fusion, attention modules, semantic mapping module. In this proposed method, integration and refinement modules work together in 3D space to fuse color and depth features at low-level, middle-level and high-level. The 3D integration module fuses the color and depth features by using a 3D convolution network rather than direct concatenation or summation. Furthermore 3D refinement module realizes a top-down fashion based fusion by integrating high-level (i.e., abstract) features into low-level (i.e., detail) features so as to retain more local details. As a result, the continuity of the semantic segmentations is improved. In addition, the

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

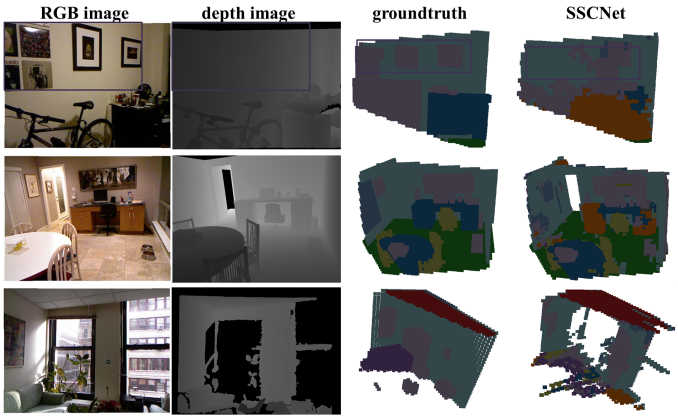

Fig. 1. In the first row, given a depth image, some objects are hard to separate, while it can be easily distinguished by using RGB information. In the second row, the semantic segmentation of SSCNet is discontinuous. In the third row, the results of SSCNet will output input-irrelative semantic label.

proposed multi-level integration and refinement strategy can eliminate such data pollution problem observed by some low-level feature fusion methods [7]. The attention module can bias input-related features so as to suppress input-irrelevant sematic labeling. Finally, our proposed method is evaluated on the NYU [8] and NYUCAD [1] datasets and experimental results have illustrated that our proposed method can outperform many state-of-the-art methods including non-deep-learning based methods, depth based methods (e.g., SSCNet and extended SSCNet), 2D color fusion methods.

II. Related Work

In this section, we discuss related works on semantic scene completion and attention mechanism.

Scene semantic completion. In 2D space, 2D-based methods [9], [10] usually treat the depth image as additional information and have achieved outstanding performances, while the depth information is crucial to 3D scene completion. In 3D space, the most methods usually convert the depth image into a volumetric representation (flipped truncated signed distance function, flip-tsdf), which can provide general geometries. The goal of scene semantic completion is completing and understanding a 3D scene. In the beginning, Zheng et al. [11] uses some pre-defined rules for 3D scene completion in the occluded parts, then contacts groups and semantic labeling. Firman et al [1] adopts random forest to make predictions for the occluded voxels and update the values of the voxels for completing the occluded 3D object shapes. These methods are a two-step process of segmenting scene and completing scenes or objects. In 2017, Song et al [2] propose a 3D CNN (called SSCNet) to formulate scene completion and semantic labeling as a joint task from a single depth. The recall of SC results of SSCNet has achieved over 90%, which almost higher over 10% than most existing methods and higher. However, it ignores local details. And due to the limitation of computational memory, the prediction of SSCNet is low-resolution, which downscale four times in each axis relative to the input's size. These result in low-quality semantic labeling. Zhang et al. [12] design a spatial group convolution to output predictions of the same size as input. Wang et al. [13] use a generative adversarial networks (GAN) to add global constraints to suppress abnormal voxels. On the other hand, many works [4]–[6] first produce 2D semantic segmentations or 2D features, then project them into 3D visible surface and fuse them into depth features for scene semantic completion. These methods have an improvement on scene semantic labeling. Some other methods [7], [14] focus on the fusion of RGB images and depth images in 3D scene space. Guedes et al. [7] assess two fusion way, including early fusion and mid-level fusion, and preliminary results are not better than SSCNet. Li et al. [14] propose a dimensional decomposition residual network to fuse more multi-level features of depth and RGB for scene semantic completion, which have extremely few improvements on SSC.

Attention Mechanism. In many papers, features in CNNs have been proved to have more high specific in the channel. CAM [15] use a new weight layer in channel at the end of the network to select task-related features. Hu et al. [16] build a SE block to extract feature descriptors to pick input-related features for the classification task, which can be inserted into everywhere of the network. In our paper, we can use channel-wise attention to select input-related features and remove unrelated features, which enhance final-feature representations.

III. Methodology

An overview of our AM^{2}FNet is given in Fig.2. Our AM^{2}FNet has six main components, including depth feature module, color feature module, 3D integration module for multi-modality feature fusion, 3D refinement module for multi-scale feature fusion, attention modules, semantic mapping module. Our AM^{2}FNet uses two module (depth feature module, color feature module) to extract features from RGB and depth image separately. And these two kinds of features are fused through other four modules. In below subsections, we firstly introduce the input of the first two modules briefly ans then explain other four modules in detail, where the integrate module and the refine module are paired together.

A. depth feature module and color feature module

The structure of depth feature module and color feature module follow previous work [2] (SSCNet). Following the most methods, flip-tsdf [2] is adopted to encode the depth image into 3D space by compute distance between each voxel with its closest voxel in visible surfaces, which can easily distinguish from the visible surfaces and the occluded parts. Meanwhile, we can project the corresponding RGB image to visible surfaces of three volumetric data respectively according to the depth image.

For the input of these two modules, the grid size of inputs is 0.02m (3D space with 4.8m horizontally and 2.88m vertically and 4.8m in depth),results in a 240x144x240 volume for the depth feature module and three 240x144x240 volumes for the color feature module.

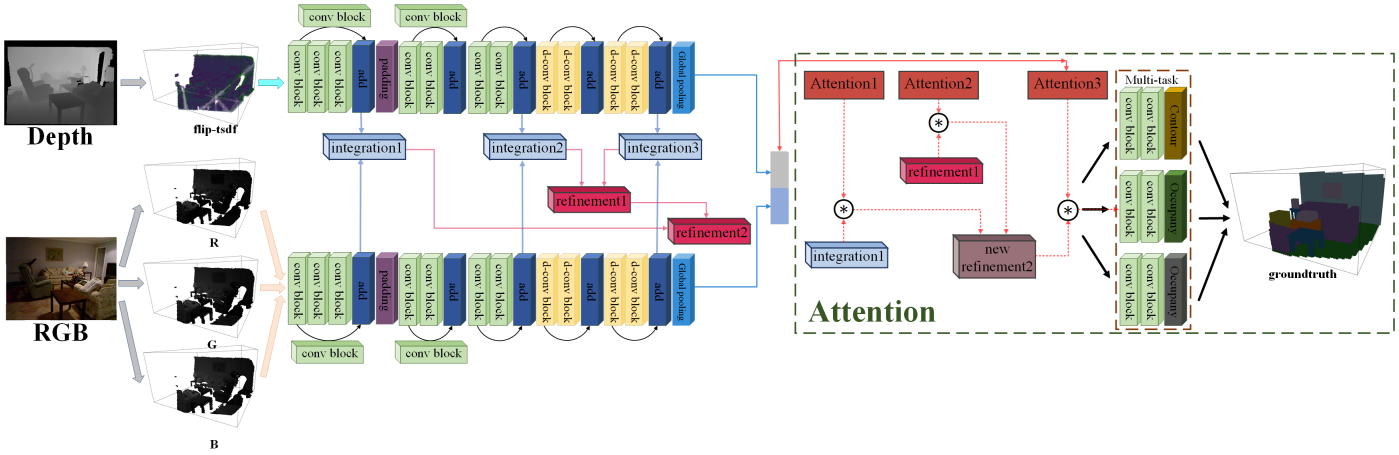

Fig. 2. The process of our proposed AM^{2}FNet work for scene semantic completion. Our network takes RGB and depth image as input, and predicts semantic labels, occupancies and contours simultaneously after multi-modality feature fusion and multiscale feature fusion. The visible surfaces of flip-tsdf are in purple. And occluded parts closed to surface are in green. The symbol ⊛ stands for channel-wise multiplication.

B. The integration module and the refinement module

The main issue of this subsection is how to effectively extract color features along with flip-tsdf (computed by depth image) and combine these two-modality features at multiply scales to produce discriminative features for ssc task. Our integration module extend RDFNet [9] based on RefineNet [17] to handle 3D multimodal feature fusion of depth and RGB and multiscale-feature fusion. The refinement module in a top-down manner is better than that in concatenation. These fusion modules have three key components: residual convolution unit (RCU), multi-resolution fusion (MRF) and chained residual pooling (CRP), as shown in Fig.3(a). The RCU consists of an adaptive convolution set, which is summation of relu-conv(3)-relu-conv(3) and the original input. MRF is used to match two different-resolution features by a 3x3x3 convolution and a upsampling operation in each pathway. CRP adopts multiple hierarchical summations by a 5x5x5 pool and a 3x3x3 convolution, which aim to extract features from a large receptive field.

1) The integration module: The purpose of the integration module is effective fuse depth features and color features, which follow MMFNet [9]. In our proposed network, we first use a 1x1x1 convolution to downscale the number of channel 4 times. Then each feature goes through two RCUs and a 3x3x3 convolution. Finally, we sum two-modal features and adopt a single-level CRP. The process is shown in Fig.3(b). The importance of each modality feature can be controlled by the learnable parameters in the convolution after RCUs.

2) The refinement module: Two integrate results go through two RCUs. Then they fused by an MMF, a two-level CPR and an another RCU finally, as shown in Fig.3(c).

C. Attention module

This section introduce our attention module. The attention module is a computational unit, which can map intermediate features $\mathbf{f} \in R^{D\times H\times W\times C}$ to same-size features $\mathbf{f}^{'} \in R^{D\times H\times W\times C}$. It can significantly reduce noise and enhance feature representations. Our design of attention module focus on the selection of the input-related features in channel. In the first step, it need to be computed as

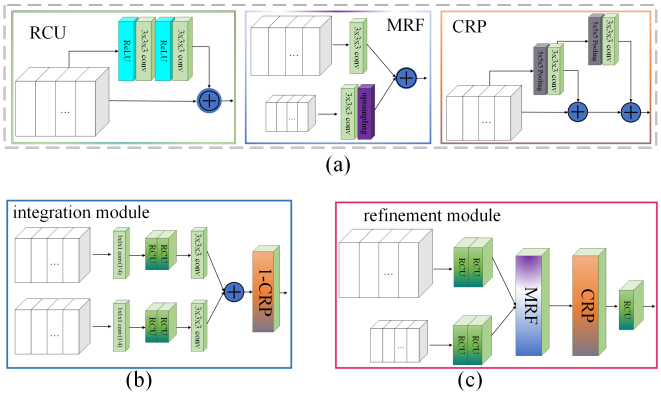

Fig. 3. The structure of the integration module and refinement module.

$$\mathbf{k} = conv\_relu(W_2 * conv\_relu(W_1 * \mathbf{f_g})) \quad (1)$$

$$\mathbf{f_g} = \{global\_pooling(\mathbf{f_d}); global\_pooling(\mathbf{f_{rgb}})\} \quad (2)$$

Here $\mathbf{f}$ is a intermediate feature descriptor. In our paper, we extract two feature descriptors of the final features ($\mathbf{f_d}$ and $\mathbf{f_{rgb}}$) in depth and RGB pathways by global average pooling and concatenate these to produce $\mathbf{f_g}$. $conv\_relu$ denotes a convolution operation with a relu operation. W_1 W_2 are the learned parameters. $\mathbf{k}$ is a vector for weighting each channel. Then it can be inserted to the correspond features, as follow:

$$\mathbf{f}^{'} = \mathbf{k} \circledast \mathbf{f} \quad (3)$$

$\circledast$ stands for channel-wise multiplication. The features in the cth channel be multiplied by a same value k_c. It also can be formulated as

$$f^{'}_{c,x,y,z} = k_c * f_{c,x,y,z} \quad (4)$$

We use attention module to produce three weighed vectors on three intermediate features, which is integration1, refinement1 and refinement2 in Fig.2 respectively.

D. semantic mapping module

Finally, our AM^{2}FNet maps the final features to the semantic labels $C = \{c_0, ..., c_{N+1}\}$, where c_0 represents the empty label. Here we also consider a channel-wise attention module for selecting input-related features in Fig. 2 to improve SSC results.

E. Multi-task Training

During the training phase, instead of treating SSC task as a single problem, we formulate it as a joint task by exploring the complementary information. The basic branch is the supervision of SSC, which is voxel-to-voxel. When we train a segmentation network, it tends to overemphasize certain parts of features. This is rooted in the downsampling path causing spatial information loss along with feature abstraction, especially in contours, while contour information can indeed provide good complementary cues for segmentation. To this end, we propose a contour branch in the end of the network with auxiliary supervision to achieve finer predictions. On the other hand, we add a two-class segmentation to increase space between empty and semantic labels. The structures of all three branches are almost consistent, which is only different in the last layers.

The parameters of depth feature module and RGB feature module (defined as W_d and W_{rgb}), integrating features (called W_{inte}) and refining features (called W_{ref}) are shared and updated for these three closely related tasks simultaneously. And the parameters of scene semantic completion branch, occupancy branch and contour branch (denoted as W_{ssc}, W_{cont}, W_{occ}) are updated independently for inferring the semantic predictions, occupied predictions and contoured predictions separately, which can increase discriminative characteristic of feature representations. And we can use occupied predictions and contoured predictions for additional post-processing in the future work.

Our AM^{2}FNet can be trained in a end-to-end manner. The loss function of our network can be formulated as

$$\begin{aligned} L_{toal}(p; W_{all}) &= L_{multi-task} + L_{reg} \\ &= -\alpha_0 \sum_{x,y,z} L_{sm} p^{ssc}_{x,y,z}(W_f, W_{ssc}), y^{ssc}_{x,y,z} \\ &\quad - \alpha_1 \sum_{x,y,z} L_{sm} p^{occ}_{x,y,z}(W_f, W_{occ}), y^{occ}_{x,y,z} \\ &\quad - \alpha_2 \sum_{x,y,z} L_{sm} p^{cont}_{x,y,z}(W_f, W_{cont}), y^{cont}_{x,y,z} \\ &\quad + \beta ||W_{all}||_2^2 \end{aligned} \tag{5}$$

where the first part is the multitask loss term (including ssc, occupancy and contours), and the latter is the regularization terms of W_{all}. (x, y, z) is the voxel position in the 3D space. $p^{ssc}_{x,y,z}(W_f, W_{ssc})$, $p^{occ}_{x,y,z}(W_f, W_{occ})$, $p^{cont}_{x,y,z}(W_f, W_{cont})$ denote ssc predictions, occupied predictions and contoured predictions respectively. y stands for groundtruth. L_{sm} is the voxel-wise softmax cross-entropy loss. t_c is class label for class c. W_f are parameters of the sharing network ($W_f = \{W_d, W_{rgb}, W_{inte}, W_{ref}\}$). $\{\alpha_0, \alpha_1, \alpha_2, \beta\}$ are hyperparameters for balancing all components of AM^{2}FNet. Actually there also exist hyper parameters for balancing the loss between different semantic labels. All parameters $W_{all} = \{W_f, W_{ssc}, W_{occ}, W_{cont}\}$ are optimized by minimizing the total loss function L_{total}.

IV. Experiments

In this section, we evaluate our proposed AM^{2}FNet on NYU [8] and NYUCAD datasets [1] with state-of-the-art methods. Both quantitative and qualitative results will be given.

A. Experimental settings

We implement our network in caffe [18]. Due to memory limitation, we set the batch size to 1 and then update our network in each four iterations. The input size in the training phase is set to 160x124x160, which is randomly cropped from a 240x144x240 volume. And in the testing phase, our network accepts the full size of volumes. We pretrain two pathway of depth and RGB firstly. And we set learning rate of this part to 0.004 and 0.01 after some iterations, while the learning rate of the rest part is 0.01. We use a SGD optimizer with a momentum of 0.9, a weight decay of 0.0005. $\{\alpha_0, \alpha_1, \alpha_2, \beta\}$ are set to 2, 1, 10, 0.2 respectively. It takes us around 2-3 days to accomplish the training phases on NVIDIA TITAN GPU(12G).

B. Datasets

NYU [8] is a real indoor scene dataset of 1449 RGB images and the corresponding depth images that are captured by Kinect, which is divided into 795 training samples and 654 testing samples. We follow [19] to attain these by voxelizing the 3D mesh annotations for ssc task. There exist some misalignments between manually labeled groundtruths and their correspond depth images. NYUCAD [1] adopts to generate depth images through projecting 3D annotations to match depth images and annotations.

C. Evaluate Metrics

As our evaluation metric, the voxel-wise precision, recall and intersection over union (IoU) are used. Following [1], we don't consider voxels outside the view and the room. For ssc task, we evaluate the IoU of each class on both visible surfaces and occluded parts, while we evaluate precision, recall and IoU of the binary predictions on occluded parts for sc task.

D. Quantitative results

Table I and Table II presents the quantitative results on NYU and NYUCAD dataset respectively with comparison to the state-of-the-art methods. In Table I (① stands for on-deep-learning based methods. ② is SSCNet. ③ is 2D color

fusion methods, while ④ is a class of methods extended SSCNet.), we compare our method with the benchmark SSCNet, which achieve higher precision and IoU of scene completion results by respectively. And AM^{2}FNet almost have the highest precision and IoU of scene completion for each class. It indicates that RGB information facilitates the improvement of semantic scene completion. Compared to the same type methods, our proposed AM^{2}FNet also achieve the highest IoUs of semantic scene completion results, which demonstrate the effectiveness of our network. It can be illustrated that 3D color features indeed improve inferring 3D geometries than 2D-based methods in term of precision and IoU on SC results. In addition, our IoU of semantic scene completion is slightly worse than ③ . The main reason is that 2D network of ③ must be trained on other bigger image datasets. The semantic scene completion mainly relies on 2D segmentation of visible surfaces. In fact, 3D geometries can help to improve semantic scene completion results and vice versa. Therefore our proposed method has more research significance.

Also, we use Table II (the sign of the type is consistent with Table I) to further demonstrate robustness and generalization of our proposed network. And the comparison results have the same tendency. We outperform SSCNet by a significant margin of 8.5% on SSC.

E. Qualitative results

To verify the effect of our AM^{2}FNet in detail, some visualization results on NYU and NYUCAD dataset are shown in Fig.4. It can be seen that our results contain less abnormal voxels, stronger structures, more accurate objects and more detailed contours. And our method can easily distinguish the objects in the wall, which is hard for SSCNet. For example, in the first column, some pictures on the wall can be recognized by using our methods, where SSCNet fails to tell them apart. Actually, the groundtruths of the dataset have some inaccurateness (e.g., pictures in the second column), which use regular groungtruths to represent irregular objects. The object outside the room is not considered in our work.

F. Ablation Study

The impact of input size. In Table III, the input size of Ours(single_1) and Our(two_1) is bigger than Ours(single) and Our(two). And SSC results have been improved. It demonstrates that our method can achieve better results with larger input size.

Is multimodal feature fusion helpful? According to the above quantitative results, the effectiveness of multimodal feature fusion has been demonstrated.

Is multiscale feature fusion helpful? In order to investigate the effect of multiscale feature fusion, we provide results of different scale fusion. In the last part of Table I, Ours(single) and Ours(two) stand for using outputs of integration3 module and refinement1 module as the input of the post-processing module. Besides, due to the limitation of memory and fair comparison, we downsample the final predictions of 120*72*120 size into 60*36*60 size. Our predictions will be more detailed with the same results.

What is the effect of Attention module? To build the relationship between global input-related descriptors with predictions, we propose to add attention vectors into the intermediate features in channel. In the last two rows of Table I and Table II, we observe that Our gives 1.8% and 2.5% improvement in IoU of semantic scene completion than Ours(no_att).

V. Conclusion

In this paper, we propose a 3D convolutional network, called AM^{2}FNet, for semantic scene completion by fusing RGB and depth features. Experimental results show AM^{2}FNet outperforms the state-of-the-art methods, including non-deep-learning based methods, depth based methods (e.g., SSCNet and extended SSCNet), 2D color fusion methods. The effectiveness of each component (including the integration module, the refinement module and attention module) is demonstrated. We also demonstrates that our method can easily achieve better results from several aspects.

References

[1] M. Firman, O. M. Aodha, S. Julier, and G. J. Brostow, "Structured prediction of unobserved voxels from a single depth image," in *Computer Vision and Pattern Recognition*, 2016.

[2] S. Song, F. Yu, A. Zeng, A. X. Chang, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Computer Vision and Pattern Recognition*, 2017.

[3] L. Zhang, L. Wang, X. Zhang, P. Shen, M. Bennamoun, G. Zhu, S. A. A. Shah, and J. Song, "Semantic scene completion with dense crf from a single depth image," *Neurocomputing*, vol. 318, pp. 182–195, 2018.

[4] Y. Guo and X. Tong, "View-volume network for semantic scene completion from a single depth image," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 726–732.

[5] S. Liu, Y. Hu, Y. Zeng, Q. Tang, B. Jin, Y. Han, and X. Li, "See and think: disentangling semantic scene completion," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2018, pp. 261–272.

[6] M. Garbade, Y.-T. Chen, J. Sawatzky, and J. Gall, "Two stream 3d semantic scene completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[7] A. B. S. Guedes, T. E. D. Campos, and A. Hilton, "Semantic scene completion combining colour and depth: preliminary experiments," 2018.

[8] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[9] S. Lee, S. J. Park, and K. S. Hong, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," in *IEEE International Conference on Computer Vision*, 2017.

[10] L. Schneider, M. Jasch, B. Fröhlich, T. Weber, U. Franke, M. Pollefeys, and M. Rätsch, "Multimodal neural networks: Rgb-d for semantic segmentation and object detection," in *Scandinavian Conference on Image Analysis*. Springer, 2017, pp. 98–109.

[11] Z. Bo, Y. Zhao, J. C. Yu, K. Ikeuchi, and S. C. Zhu, "Beyond point clouds: Scene understanding by reasoning geometry and physics," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[12] J. Zhang, H. Zhao, A. Yao, Y. Chen, L. Zhang, and H. Liao, "Efficient semantic scene completion network with spatial group convolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 733–749.

[13] Y. Wang, D. J. Tan, N. Navab, and F. Tombari, "Adversarial semantic scene completion from a single depth image," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 426–434.

TABLE I
SSC RESULTS ON NYU DATASET. THE HIGHEST VALUES ARE IN BOLD.

		scene completion			scene semantic completion											
Type	Methods(train)	pre.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
①	Lin et al(NYU)	58.5	49.9	36.4	0.0	11.7	13.3	14.1	9.4	29.0	24.0	6.0	7.0	16.2	1.1	12.0
	Geiger(NYU)	65.7	58.0	44.4	10.2	62.5	19.1	5.8	8.5	40.6	27.7	7.0	6.0	22.6	5.9	19.6
②	SSCNet(SUNCG+NYU)	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(NYU)	57.0	**94.5**	55.4	15.1	94.7	24.4	0	12.6	32.1	35	13	7.8	27.1	10.1	24.7
	SSCNet*(NYU)	60.9	91.5	57.5	18.9	92.7	24.4	8.2	17.4	52.1	42.7	16.0	13.9	33.5	12.1	30.2
③	two-stream(pre_trained+NYU)	-	-	60	9.7	93.4	25.5	21	17.4	55.9	49.2	17	27.5	39.4	19.3	34.1
	VVNet(pre_trained+SUNCG+NYU)	69.8	83.1	61.1	19.3	94.8	28	12.2	19.6	57	50.5	17.6	11.9	35.6	15.3	32.9
	SATNet(pre_trained+SUNCG+NYU)	67.3	85.8	60.6	17.3	92.1	28	16.6	19.3	57.5	53.8	17.7	18.5	38.4	18.9	**34.4**
④	DDRNet(NYU)	71.5	80.8	61	21.1	92.2	33.5	6.8	14.8	48.3	42.3	13.2	13.9	35.3	13.2	30.4
	DCRF(NYU)	-	-	-	18.1	92.6	27.1	10.8	18.8	54.3	47.9	17.1	15.1	34.7	13	31.8
⑤	Ours(single)	67.6	85.2	60.4	16.2	88.6	26.5	11.2	18.6	49.8	41.2	15.3	15.4	34.7	14.2	30.2
	Ours(two)	67.5	84.7	60.0	15.9	92.3	26.6	11.4	17.9	52.3	46.1	16.3	18.0	34.7	13.8	31.4
	Ours(no_att)	70.9	82.3	61.2	19.1	92.0	26.0	10.6	18.8	52.1	47.1	16.9	9.5	35.8	13.7	31.1
	Ours	**72.1**	80.4	**61.3**	19.3	92.6	26.1	11.1	19.1	51.9	47.0	16.7	14.9	35.9	14.0	31.7

TABLE II
SSC RESULTS ON NYUCAD DATASET. THE HIGHEST VALUES ARE IN BOLD.

		scene completion			scene semantic completion											
Type	Methods(train)	pre.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
①	Zheng et al(NYU)	60.1	46.7	34.6	-	-	-	-	-	-	-	-	-	-	-	-
	Firman et al(NYU)	66.5	69.7	50.8	-	-	-	-	-	-	-	-	-	-	-	-
②	SSCNet(SUNCG+NYU)	75.0	**96.0**	73	-	-	-	-	-	-	-	-	-	-	-	-
	SSCNet(NYU)	75	92.3	70.3	-	-	-	-	-	-	-	-	-	-	-	-
	SSCNet*(NYU)	76.8	95.4	74	37.1	92.3	47.8	8	37.4	60.9	59.8	32	9.2	44.4	23.1	41.1
③	two-stream(pre_trained+NYU)	-	-	76.1	25.9	93.8	45.9	33.4	31.2	66.1	56.4	31.6	38.5	51.4	30.8	**46.2**
	VVNet(pre_trained+SUNCG+NYU)	86.4	92	**80.3**	-	-	-	-	-	-	-	-	-	-	-	-
④	DDRNet(NYU)	**88.7**	88.5	79.4	54.1	91.5	56.4	14.9	37	55.7	51	28.8	9.2	44.1	27.8	42.8
	DCRF(NYU)	-	-	-	35.5	92.6	52.4	10.7	40	60	62.5	34	9.4	49	26.5	43
⑤	Ours(no_att)	85.6	92.1	79.6	36.9	92.2	53.1	15.5	38.6	63.1	60.0	32.5	11.5	50.1	25.3	43.5
	Ours	87.2	91.0	80.2	39.0	92.1	54.2	17.6	43.6	64.1	58.6	33.2	12.3	50.0	25.8	44.6

TABLE III
IMPACT OF THE INPUT SIZE FOR SSC RESULTS ON NYU DATASET.

	scene completion			scene semantic completion											
Methods(train)	pre.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
Ours(single)	67.6	85.2	60.4	16.2	88.6	26.5	11.2	18.6	49.8	41.2	15.3	15.4	34.7	14.2	30.2
Ours(single_1)	66.5	85.8	59.7	18.7	93.4	26.2	7.9	18.2	54.1	47.7	16.9	17.3	35.9	15.2	32.0
Ours(two)	67.5	84.7	60.0	15.9	92.3	26.6	11.4	17.9	52.3	46.1	16.3	18.0	34.7	13.8	31.4
Ours(two_1)	67.7	85.0	60.3	15.9	93.1	26.8	11.0	20.2	55.4	49.8	17.5	17.4	36.1	15.2	32.6

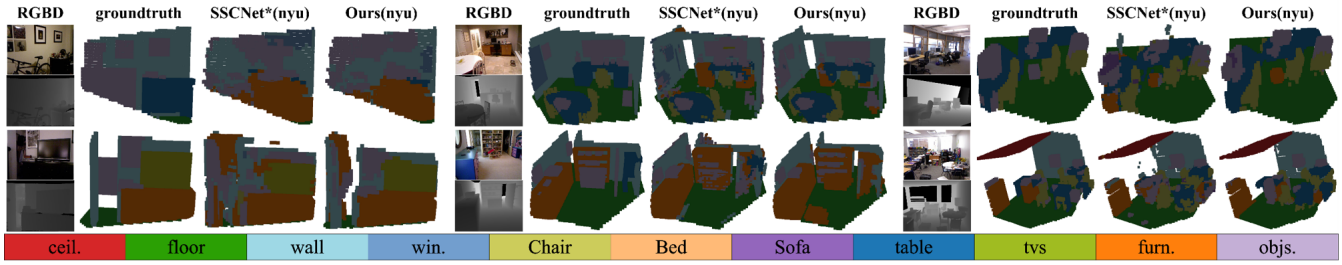

Fig. 4. Qualitative results on NYU and NYUCAD datasets. The data in the first row comes from NYU dataset. The data in rest come from NYUCAD dataset.

[14] J. Li, Y. Liu, D. Gong, Q. Shi, X. Yuan, C. Zhao, and I. Reid, "Rgbd based dimensional decomposition residual network for 3d semantic scene completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7693–7702.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[19] R. Guo, C. Zou, and D. Hoiem, "Predicting complete 3d models of indoor scenes," *Computer Science*, 2017.