

# Attention-Based Multi-Modal Fusion Network for Semantic Scene Completion

Siqi Li,<sup>1</sup> Changqing Zou,<sup>2</sup> Yipeng Li,<sup>3</sup> Xibin Zhao,<sup>1</sup> Yue Gao<sup>1\*</sup>

<sup>1</sup>BNRist, KLISS, School of Software, Tsinghua University, China

<sup>2</sup>Huawei Noah's Ark Lab, <sup>3</sup>Department of Automation, Tsinghua University, China  
lsq19@mails.tsinghua.edu.cn, aaronzou1125@gmail.com, {liep, zxb, gaoyue}@tsinghua.edu.cn

## Abstract

This paper presents an end-to-end 3D convolutional network named attention-based multi-modal fusion network (AMFNet) for the semantic scene completion (SSC) task of inferring the occupancy and semantic labels of a volumetric 3D scene from single-view RGB-D images. Compared with previous methods which use only the semantic features extracted from RGB-D images, the proposed AMFNet learns to perform effective 3D scene completion and semantic segmentation simultaneously via leveraging the experience of inferring 2D semantic segmentation from RGB-D images as well as the reliable depth cues in spatial dimension. It is achieved by employing a multi-modal fusion architecture boosted from 2D semantic segmentation and a 3D semantic completion network empowered by residual attention blocks. We validate our method on both the synthetic SUNCG-RGBD dataset and the real NYUv2 dataset and the results show that our method respectively achieves the gains of 2.5% and 2.6% on the synthetic SUNCG-RGBD dataset and the real NYUv2 dataset against the state-of-the-art method.

## Introduction

Understanding and reconstructing a 3D scene from partial observations is a very important technique which has received increasing research attention in recent years due to its commercial potential in a large variety of robotics and vision tasks such as robotic navigation (Gupta, Arbelaez, and Malik 2013), autonomous driving (Laugier et al. 2011), and scene reconstruction (Hays and Efros 2007; Han et al. 2019). Given a single depth image or RGB-D images of a 3D scene, many papers (Gupta, Arbelaez, and Malik 2013; Ren, Bo, and Fox 2012; Firman et al. 2016) have been proposed to complete or segment the 3D scene with neural networks. More recently, a set of methods (Song et al. 2017; Liu et al. 2018; Guo and Tong 2018; Zhang et al. 2018; Li et al. 2019b) have been developed to automatically predict the semantic labels, together with completing the 3D geometry, of the objects in a 3D scene from a single view of the scene using convolutional neural networks.

\*Corresponding author

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

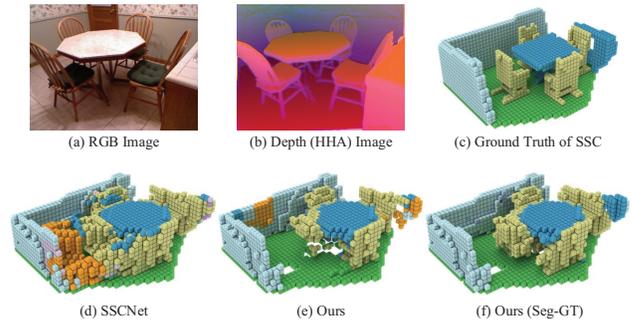


Figure 1: Given RGB-D images (a-b) as input, the proposed AMFNet can produce more accurate scene completion and scene segmentation result (e) than the previous methods (e.g., SSCNet (Song et al. 2017) (d)). Directly boosting the AMFNet from the 2D segmentation ground truth of the input RGB-D images can even produce a result closer to the ground truth (c).

This line of literature focusing on the task of semantic scene completion (SSC) was initiated by those methods (Song et al. 2017; Zhang et al. 2018) which take as input only the depth information. Several later works (Liu et al. 2018; Li et al. 2019b; Guedes, de Campos, and Hilton 2018) argue that the RGB color information, which is often captured together with the depth information by an RGB-D sensor, could be used to improve the performance of the networks which take only the depth information. The experimental results from these works reveal that the RGB color channels provide complementary information to the depth information and higher accuracy of SSC can be achieved by using the fused information.

In this work, we propose a novel approach for the task of semantic scene completion, which makes use of not only the information of RGB color channels and depth but also the “experience” of deep network learning 2D semantic segmentation. Specifically, we train a model that leverages 2D semantic segmentation information to guide both 3D completion and semantic labeling (semantic segmenta-

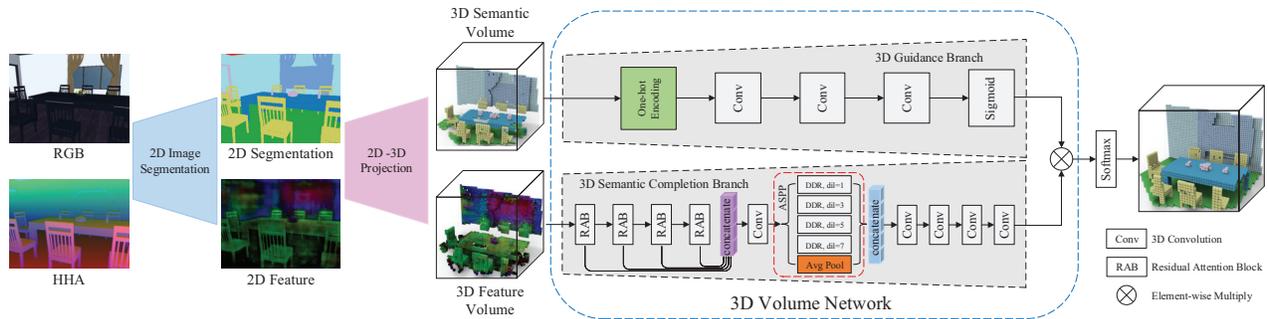


Figure 2: Architecture of AMFNet. Taking RGB-D images (separated to a RGB and a HHA image) as input, AMFNet predicts voxel occupancy and object labels of the scene simultaneously. It boosts the 3D completion and segmentation from an initial 3D semantic feature volume produced by computing the 2D-3D projection of the results of a 2D segmentation network.

tion) in the SSC task. Some prior works (Chen et al. 2018; Yang et al. 2018; Li et al. 2019a; Xiong et al. 2019) show that 2D segmentation is more accurate than the 3D SSC task. Our model is based on the assumption that the image features which are reliable for 2D semantic segmentation should also be reliable for semantic scene completion. With this assumption, we propose a novel two-branch multi-modal fusion network where the 3D segmentation information, which is boosted from 2D segmentation, could guide both 3D completion and semantic labeling feasibly. Moreover, due to the inherent sparsity of 3D data (most voxels in a 3D scene are empty), we believe that some features (e.g., color and texture) associated with the empty voxels are valueless. Therefore, we propose a residual attention block (RAB) based 3D network to make full use of reliable depth cues in the spatial dimension for semantic scene completion.

The proposed network contains three sequential modules: a 2D segmentation module which extracts 2D image features and produces 2D semantic labels, a 2D-3D projection layer which generates a 3D semantic feature volume from the output of the 2D segmentation module, and an attention-driven two-branch 3D volume network (named 3D volume network) which infers a complete 3D volume with semantic labels from the initial 3D scene.

We validate our approach on both the synthetic SUNCG-RGBD dataset (Liu et al. 2018) and the real NYUv2 dataset (Silberman et al. 2012), and the results show that our method achieves a gain of 2.5% on the SUNCG-RGBD dataset and a gain of 2.6% on the NYUv2 dataset against the state-of-the-art method (Li et al. 2019b). Our analytical study also observes that directly boosting the AMFNet from the 2D segmentation ground truth of the input RGB-D images (also see Figure 1(f)) can achieve a gain of 3.8% on the synthetic SUNCG-RGBD dataset and a gain of 4.3% on the real NYUv2 dataset against the state-of-the-art method (Li et al. 2019b).

Our main contributions can be summarized as follows:

- A multi-modal fusion network boosted from 2D semantic segmentation is proposed for semantic scene completion. Compared to previous works (Song et al. 2017; Liu et al. 2018; Li et al. 2019b), the major advantage of

this network is that 2D semantic segmentation is used to guide and improve the 3D feature extraction for the SSC task. We demonstrate this advantage can significantly boost the overall performance of the task of SSC by improving both scene completion and segmentation.

- A residual attention block (RAB) based network is proposed for scene semantic completion. Our experiments show that the proposed RAB is particularly effective for some particular objects in a 3D scene.
- The proposed end-to-end framework achieves state-of-the-art performance on two large-scale datasets, i.e., the NYUv2 (Silberman et al. 2012) and the SUNCG-RGBD (Liu et al. 2018) datasets.

## Related Works

**Semantic Scene Completion** The pioneering work for the task of semantic scene completion was proposed by Song et al. in (Song et al. 2017). Although this method achieves good performance in terms of performing both scene completion and semantic segmentation simultaneously from a single-view depth image within a task, it requires an extremely time-consuming data pre-processing step of computing the flipped truncated signed distance function (fTSDF) used for eliminating intense gradients and view dependency.

Motivated by the fact that most of the voxels in the 3D volume are empty and useless, Zhang et al. further improved the SSCNet by applying a spatial group convolution (SGC) in EsscNet (Zhang et al. 2018). This method divides the input fTSDF into different groups and then forwards them to a 3D sub-manifold sparse convolutional network (Graham, Engelcke, and van der Maaten 2018). A limitation of the EsscNet is that the input features are empirically divided into several groups, which might lead to performance degradation.

More close to our work, SATNet (Liu et al. 2018) is a model that performs the SSC task from 2D semantic features. This model contains three sequential modules similar to the proposed AMFNet: a 2D network to extract semantic features from RGB-D images, a 2D-3D projection layer, and

a 3D network to complete the scene. Generally, the SATNet has a data processing flow (i.e., boosting 3D completion and segmentation from the 2D-3D projection of the 2D semantic features) similar to the 3D-semantic completion branch (the bottom one in Figure 2) of AMFNet. DDRNet (Li et al. 2019b) is a light-weight deep model with an architecture similar to SATNet, of which the major contribution is a dimensional decomposition residual (DDR) block, which reduces the network parameters dramatically by decomposing the traditional 3D convolution block into consecutive layers channel-wise. The proposed AMFNet follows data processing flow similar to SATNet and DDRNet, but rather than perform 3D scene completion and semantic segmentation in a single branch, AMFNet employs two separated branches for 3D scene completion and semantic segmentation respectively, which enables the 2D semantic segmentation information to be used to explicitly guide the 3D scene completion and semantic segmentation simultaneously.

**Attention Mechanism** A large number of CNN-based methods (Wang et al. 2017; Hu, Shen, and Sun 2018) use the attention mechanism to improve the performance of the classification task. Hu et al. (Hu, Shen, and Sun 2018) employ an attention scheme which infers the importance of multiple channels for image classification. This method achieves significant gains in classification accuracy by making the model focus on meaningful regions of the image and ignores meaningless regions with the attention scheme. Woo et al. (Woo et al. 2018) propose an attention scheme named convolutional bottleneck attention module (CBAM) to estimate channel-wise attention weight and spatial-wise attention weight. Inspired by the great success of the application of attention mechanism in the image classification task and the inherent sparsity of 3D data, we believe that the ability of the attention mechanism, which makes the model focus on important parts, maybe also helpful for the task of SSC. In our network, to inject the attention into the 3D volume network without significant parameter increase, we graft the CBAM attention module onto a computation-efficient residual block to form a novel attention-injected residual block (RAB) and use this RAB for 3D feature extraction.

## Method

In this section, we first present the structure of the proposed AMFNet and then detail each module.

The proposed AMFNet, as illustrated in Figure 2, mainly contains three sequential modules: a 2D segmentation network, a 2D-3D projection layer, and a two-branch 3D volume network. This network takes single-view RGB-D images as input and outputs occupancy and semantic labels for all voxels in the scene. The whole network can be trained in an end-to-end manner. We next introduce each module in the sequence of the flow of data processing.

### 2D Segmentation Network

The 2D segmentation network extracts 2D geometry features and performs 2D semantic segmentation from the input

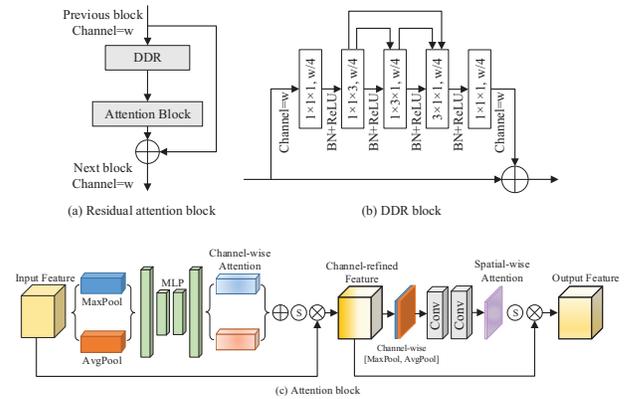


Figure 3: Illustration of the proposed residual attention block (RAB). RAB has a structure similar to the DDR block (Li et al. 2019b) but with both channel-wise and spatial-wise attention injected.

RGB-D images. Since it is unlikely that a CNN would automatically learn to compute the properties of geocentric pose that emphasize complementary discontinuities in the RGB-D image (e.g., depth, surface normal and height), we first compute a three-channel HHA (Gupta et al. 2014) encoding result from the input depth image. After that, the RGB image and HHA image are fed into two 2D segmentation networks that have the same structure but do not share weights, and their individual outputs are fused together by a convolution layer with a kernel size of 1. In our implementation, the 2D segmentation network has an encoder-decoder structure where the encoder uses ResNet-101 (He et al. 2016), and the decoder contains a series of up-sampling convolutions and an atrous spatial pyramid pooling (ASPP) module (Chen et al. 2016; 2018) which can extract multi-scale features. The output of the 2D segmentation network is a feature map with 64 channels and a semantic segmentation map.

### 2D-3D Projection Layer

We employ a projection layer similar to those in SATNet (Liu et al. 2018) and DDRNet (Li et al. 2019b) to generate an initial 3D semantic feature volume (can be separated into a 3D semantic volume and a 3D feature volume, as shown in Figure 2). For each input RGB-D image, given the intrinsic and extrinsic camera matrix, a 3D volume can be directly computed from the depth information (see SATNet (Liu et al. 2018) for reference). Since each pixel in the RGB-D images correspond to a tensor in the 2D feature map as well as a semantic label in the 2D segmentation map, every feature tensor and semantic label can be projected into the 3D volume at the location with the same depth value. In this way, we obtain a view-independent 3D semantic feature volume and a 3D semantic volume (only visible voxels have semantic labels), as shown in Figure 2. The 2D-3D projection operation can be recognized as reshaping the feature maps during the forward propagation. During back propagation, we only map the gradient of the voxels on the visible

surface (discarding the gradient of other voxels) to the 2D pixels and continue the back propagation in the 2D network.

### 3D Volume Network

The 3D volume network, which takes the output of the 2D-3D projection layer as input, contains two branches: one for 3D guidance information and the other for 3D semantic completion. At the end of the two branches, the outputs of the guidance branch and semantic completion branch are element-wise multiplied and forwarded into a softmax layer to generate a 3D volume with semantic labels.

**3D-Guidance Branch** The 3D-guidance branch is used to provide the guidance information for the branch of 3D semantic completion, which is boosted from an initial 3D semantic volumetric scene where visible voxels have initial semantic labels. The initial 3D semantic volume is first encoded by a one-hot encoder to achieve an ROI region (3D bounding box) for a specific category from the initial 3D semantic volume. For a dataset containing  $N$  different categories, the 3D semantic volume could be encoded by a volume with  $N$  channels. Each channel of the volume can be used to represent the ROI region of a specific category. For each channel, the value of the voxels within the bounding box of the corresponding category is set to one, otherwise, it is set to zero. The one-hot encoding introduces spatial boundary constraints into the network for each category, which improves the prediction of 3D semantic volume. (See the experimental analysis in Sec. for more details.) The encoded volume is then fed into three 3D dense convolution layers and a sigmoid layer to obtain the output features of the 3D-guidance branch. It is worth mentioning that we do not perform the one-hot encoding with instance-level semantic information which may provide more accurate boundary constraints because both datasets in our experiment do not provide instance-level segmentation information.

**3D Semantic Completion Branch** The 3D-semantic completion branch, which takes the initial 3D feature volume as input, is mainly used to infer the voxel occupancy of the 3D scene. The data processing flow is illustrated in Figure 2. The input 3D features are firstly forwarded to four RAB modules, and the outputs of each block are concatenated and forwarded to a 3D convolution layer with a kernel size of 1 for multi-level feature fusion. After that, the fused feature is further forwarded to a 3D ASPP block (i.e., multiple parallel atrous convolutional layers with different dilatation rates, which is a powerful way to handle objects with various sizes) to exploit multi-scale features. The output of the ASPP block is lastly forwarded to four cascaded 3D convolution layers to generate high-level features for the fusion with the 3D-Guidance branch. In this branch, the attention focusing on reliable depth cues in the spatial dimension are mainly captured by the RAB modules. Our experiments in the next section will show the attention scheme contributes much to the overall performance (3% gains on average on both the synthetic SUNCG-RGBD and the real NYUv2 dataset on the metric of semantic scene completion). We next detail the RAB block.

As shown in Figure 3, our RAB block consists of two sequential blocks: a DDR block and an attention block. Moreover, a shortcut connection is used to achieve the effect of the residual block (He et al. 2016). Our RAB can be functionally formulated as:

$$y = A(D(x)) + x \quad (1)$$

where  $x$  is input,  $y$  is output, and  $D$  and  $A$  denote the DDR block and attention block, respectively. Instead of the 3D-ResNet blocks, we utilize the DDR block proposed in (Li et al. 2019b), as shown in Figure 3 (b), to reduce the model parameters. The DDR block decomposes a 3D convolution with a kernel size of  $3 \times 3 \times 3$  into three consecutive layers with kernel sizes of  $1 \times 1 \times 3$ ,  $1 \times 3 \times 1$  and  $3 \times 1 \times 1$  respectively, so that the parameters could be reduced significantly.

The attention block used in RAB is a 3D adaption of the 2D attention block originally proposed in (Woo et al. 2018). It sequentially captures channel-wise attention and spatial-wise attention. The former takes into account the channel importance of the 3D features while the latter infers the spatial importance of the 3D features. Specifically, in our attention block, the 3D features of each channel are firstly aggregated by using average pooling and max pooling to generate two descriptors. Then the aggregated descriptors are forwarded to a multi-layer perceptron (MLP) to extract hidden features. The outputs of the MLP are element-wise added as the channel-wise attention. The generated channel-wise attention is multiplied by the 3D features to generate the channel-refined 3D features. In practice, the MLP has two hidden layers, and we choose 1/8 of the number of input feature channels as the hidden layer size. The spatial-wise attention is achieved via the following steps: the channel-refined 3D features are first aggregated by channel-wise average pooling and max pooling, and then the generated descriptors are forwarded to two convolution layers (the kernel sizes are  $5 \times 5 \times 5$  and  $1 \times 1 \times 1$ , respectively) to generate the spatial-wise attention. In the last stage of RAB, the input 3D features are multiplied by the attention weight to achieve the 3D features with both channel-wise and spatial-wise attention injected.

### Implementation Details

We follow (Song et al. 2017) to process the data and compute the 2D-3D projection mapping for each sample in advance following SATNet (Liu et al. 2018). The training procedure consists of two steps. We first pre-train the 2D segmentation network with the supervision of 2D semantic segmentation ground truth, and then train the whole model end-to-end. We use cross-entropy loss and an SGD optimizer with a momentum of 0.9, a weight decay of  $5e-4$ , and a batch size of 1. The learning rate of the 2D segmentation network and 3D scene completion network is 0.001 and 0.01, respectively.

## Experiments

In this section, the proposed method is evaluated and compared with the state-of-the-art methods on two public datasets, i.e., the NYUv2 dataset (Silberman et al. 2012) and the SUNCG-RGBD dataset (Liu et al. 2018).

Method	scene completion			semantic scene completion											
	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tv	furn.	objs.	avg.
Lin et al. (2013)	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 et al. (2015)	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
SSCNet (2017)	57.0	<b>94.5</b>	55.1	15.1	<b>94.7</b>	24.4	0.0	12.6	32.1	35.0	13.0	7.8	27.1	10.1	24.7
EsseNet (2018)	<b>71.9</b>	71.9	56.2	17.5	75.4	25.8	6.7	15.3	53.8	42.4	11.2	0	33.4	11.8	26.7
DDRNet (2019b)	71.5	80.8	<b>61.0</b>	21.1	92.2	<b>33.5</b>	6.8	14.8	48.3	42.3	13.2	<b>13.9</b>	35.3	13.2	30.4
Ours	67.9	82.3	59.0	16.7	89.2	27.3	<b>19.2</b>	<b>20.2</b>	<b>56.1</b>	<b>50.4</b>	<b>15.1</b>	13.5	<b>36.8</b>	18.0	<b>33.0</b>
Ours (w/o-Attn)	64.5	86.5	58.6	<b>21.3</b>	90.3	26.1	7.7	18.0	53.8	48.4	13.0	0	36.7	16.3	30.1
Ours (w/o-Seg)	68.5	78.6	57.3	13.7	<b>94.7</b>	26.9	14.8	12.7	48.5	39.8	10.4	12.2	36.2	<b>19.2</b>	29.9
Ours (basic)	64.0	86.6	58.0	20.9	94.1	27.1	15.7	12.4	46.0	45.1	13.9	0.3	32.3	15.7	29.4
Ours (Seg-GT)	66.3	80.5	57.2	20.0	78.7	27.3	20.5	21.8	56.5	53.9	19.5	18.8	40.1	19.5	34.2

Table 1: Results on the NYUv2 dataset. Bold numbers represent the best scores.

Method	scene completion			semantic scene completion											
	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tv	furn.	objs.	avg.
SSCNet (2017)	43.5	90.7	41.5	64.9	60.1	<b>57.6</b>	25.2	25.5	40.4	37.9	23.1	29.8	<b>45.7</b>	4.7	37.7
SATNet (2018)	56.7	91.7	53.9	65.5	60.7	50.3	56.4	26.1	<b>47.3</b>	43.7	30.6	37.2	44.9	<b>30.0</b>	44.8
Ours	<b>57.5</b>	91.6	<b>54.5</b>	<b>80.4</b>	<b>69.1</b>	55.0	<b>60.4</b>	<b>27.0</b>	42.2	46.7	<b>32.5</b>	42.3	36.9	27.4	<b>47.3</b>
Ours (w/o-Attn)	54.8	94.7	53.2	79.6	66.9	51.7	60.2	26.5	38.2	45.5	24.5	27.9	38.1	28.7	44.3
Ours (w/o-Seg)	54.1	94.0	52.4	76.6	61.0	53.7	54.7	26.2	37.3	<b>49.4</b>	31.8	43.7	40.7	25.0	45.5
Ours (basic)	47.5	<b>96.4</b>	46.7	71.0	49.7	41.4	57.3	21.6	41.3	46.1	30.1	<b>44.8</b>	39.9	21.4	42.2
Ours (Seg-GT)	60.6	89.1	56.3	81.3	68.5	54.1	61.8	30.2	45.9	50.7	34.3	42.7	41.9	28.4	49.1

Table 2: Results on the SUNCG-RGBD dataset (we do not compare our method to EsseNet and DDRNet because neither of them report their performances on this dataset). Bold numbers represent the best scores.

## Datasets and Metrics

**Datasets** We evaluate our method on the NYUv2 and the SUNCG-RGBD datasets. The NYUv2 (Silberman et al. 2012) is a real scene dataset, consisting of 1449 indoor scenes. The dataset is divided into 795 training and 654 testing samples, each scene associated with RGB-D images. We obtain the ground truth of the SSC task by following (Song et al. 2017). NYUv2 is a challenging dataset due to the complexity of the indoor scene and the measurement errors in the depth images caused by the Kinect data collection. SUNCG-RGBD, a synthetic dataset proposed by Liu et al. (Liu et al. 2018), is a subset of the SUNCG dataset (Song et al. 2017). It consists of 13011 training samples and 499 testing samples.

**Metrics** We mainly validate the proposed framework on two tasks, scene completion and semantic scene completion, as previous methods. We use the voxel-level intersection over union (IoU) between ground truth labels and predicted labels as the evaluation metric on both tasks. Specifically, for the task of semantic scene completion, we evaluate the IoU of each category on both the observed and occluded voxels. For the task of scene completion, we treat all non-empty voxels as the same category and evaluate the IoU of the binary predictions on the occluded voxels.

## Quantitative Comparison

**Comparison on NYUv2 dataset** Table 1 presents the comparison results of both the scene completion and the semantic scene completion on the NYUv2 dataset. Compared with previous methods, our model achieves state-of-the-art

performance on the task of semantic scene completion and ranks second on the task of scene completion. Compared to DDRNet (Li et al. 2019b), which also uses RGB-D images as input, our method achieves gains of 2.6% on the task of semantic scene completion. The quantitative result also reveals that the proposed method demonstrates superior performance against previous methods on the categories with small physical size such as *chair* (a gain of 4.9% is achieved) and the categories with variational depth information such as *window* (a gain of 5.1% is achieved). We attribute this improvement to the application of the attention block and the guidance branch in the 3D network. The attention block can make the model focus on significant parts. Therefore, the features of small objects can be acquired more effectively and further achieve better semantic scene completion performance. Meanwhile, the guidance branch, which is boosted with the sophisticated 2D segmentation network, can improve the task of 3D semantic scene completion by providing reliable guidance information for the 3D semantic scene completion branch. In the discussion section, we will validate this explanation by ablation study.

**Comparison on SUNCG-RGBD dataset** Table 2 presents the performance of our method on the SUNCG-RGBD dataset with comparison to some previous approaches. Compared with these baseline methods, our model achieves the best performance on both the tasks of scene completion and semantic scene completion. Specifically, our method outperforms the previous SSCNet by significant performance margin, which are 9.6% gains on the semantic scene completion task and 13.0% gains on the scene completion task. Compared to another recent

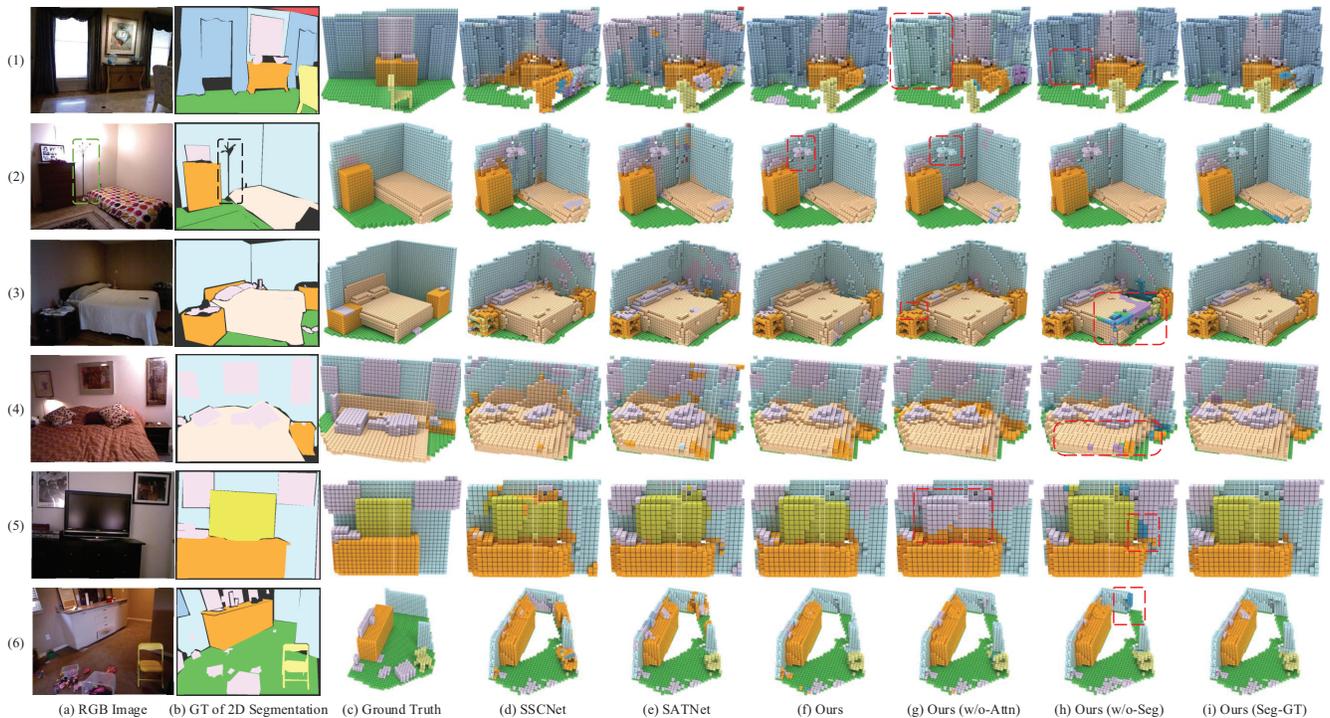


Figure 4: Qualitative results on the NYUv2 dataset. From left to right: RGB image, ground truth of 2D segmentation result, ground truth of SSC task, results generated by SSCNet (Song et al. 2017), SATNet (Liu et al. 2018), our approach, our approach with attention block removed, our approach with 3D-guidance branch removed, and our approach with the result of the 2D semantic segmentation module (see Figure 2) replaced by the ground truth. See Section 4.3 for the detailed analysis.

model SATNet, our method still has a distinct advantage (2.5% gains on semantic scene completion and 0.6% gains on scene completion). Consistent with the observations on the NYUv2 dataset, our model performs well on those categories like *chair* and *window* and respectively achieves gains of 0.9% and 4% against SATNet.

### Qualitative Analysis

Figure 4 visualizes the qualitative results of the semantic scene completion task generated by the proposed method and two previous methods, SSCNet and SATNet, on a set of representative samples from the NYUv2 dataset. It can be easily seen that our method has achieved better performance than SSCNet and SATNet. For example, our method produced much better results for objects with relatively large physical size, such as the wall and floor, in terms of segmentation consistency, as shown in row (1) and row (4). Similarly, our method can achieve a clearer 3D shape boundary of small objects, such as *chair* in row (1) and *obj.* in row (3). We attribute this to the guidance branch, which can provide the boundary constraints. In row (2), it can be seen from the RGB image of the scene that there is a lamp within the green dashed-line box, but it is regarded as background in the ground truth of both 2D segmentation and 3D semantic scene completion. Our model can reconstruct it well as

shown in column (f) (see the red dashed-line box). This indicates that our model is capable of recognizing and reconstructing objects with small physical size in the scene.

## Discussion

### How useful is the 3D-guidance branch?

To investigate if the 3D-guidance branch can be helpful for the semantic scene completion task, we study the performance of our method without the 3D-guidance branch. Specifically, we remove the 3D-guidance branch in Figure 2 and leave only the 3D-semantic completion branch. In Table 1 and Table 2, we show the quantitative performance of our model without the 3D-guidance branch (denoted as Ours (w/o-Seg)). Its performance on average for 3D semantic scene completion is 45.5% on the SUNCG-RGBD dataset, and 29.9% on the NYUv2 dataset. In contrast, the model with the guidance branch can achieve 1.8% and 3.1% performance improvements against the one with the 3D-guidance branch removed on the two datasets, respectively. Compared to the basic model (denoted as Ours (basic)) where both the attention block and the guidance branch are removed from the proposed architecture, the 3D-guidance branch can respectively improve the performance by 2.1% and 0.7% on the SUNCG-RGBD dataset and the NYUv2 dataset.

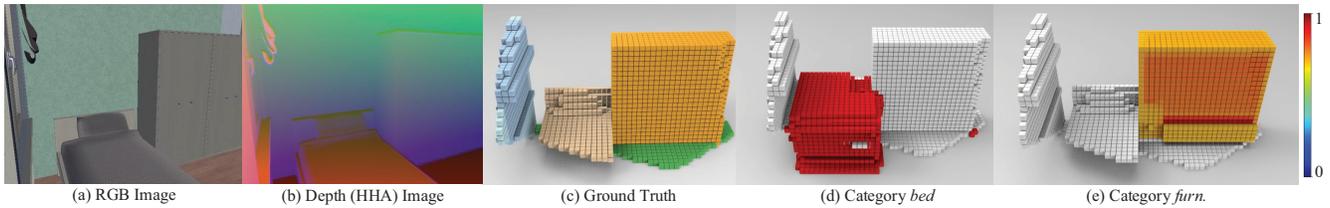


Figure 5: Visualization of the output of 3D-guidance branch. (a-b) The RGB-D images. (c) Ground truth of SSC. (d-e) 3D segmentation results of category *bed* and *furn.* (visualized from the output of the 3D-guidance branch). The color bar, which quantifies the category probability, is shown on the right as a reference. It can be seen that the output of the 3D-guidance branch does provide reliable guidance information for the task of SSC.

In Figure 5, we visualize the results of the outputs of the 3D-guidance branch. We can see that the 3D-guidance branch can accurately focus on the corresponding 3D regions and provide reliable constraints of object boundaries, which makes the 3D semantic scene completion network able to determine the accurate voxel occupancy. Therefore it benefits both the scene completion and semantic scene completion task. In Figure 4, we visually demonstrate the effect of the 3D-guidance branch. In row (3) and row (5), the model with 3D-guidance branch removed is not able to maintain the integrity of the *bed* and *wall*, as shown in column (h). However, with the 3D-guidance branch added, the proposed method can address this problem, as shown in column (f).

### How useful are the residual attention blocks?

Previous works have proved that the attention mechanism can greatly help the 2D image classification task. We here investigate if the attention mechanism can benefit the 3D semantic scene completion task. To achieve this purpose, we evaluate the performance of the proposed architecture with the attention block removed (i.e., only the DDR block in Figure 3 (a) remains). In Table 1 and Table 2, we report the performance of our model with attention mechanism removed (denoted as Ours (w/o-Attn)). The average performance of this model is 44.3% on the SUNCG-RGBD dataset and 30.1% on the NYUv2 dataset. Conversely, the model with the attention mechanism added can achieve 3% and 2.9% performance improvements on the two datasets, respectively. Compared with the basic model (denoted as Ours (basic)), the attention block can improve the performance 0.5% and 3.3% on the two datasets, respectively.

We guess this performance improvement is probably because the proposed RAB calculates both spatial-wise and channel-wise weights and is able to make the network focus on meaningful parts instead of empty voxels. Moreover, the attention mechanism in this work can improve the performance of the network on small objects, such as chairs and tables. For example, in row (3) of Figure 4, it can be seen from the RGB image that there is a pile of paper on the cabinet. The proposed method can recognize and reconstruct it while the one with attention block removed cannot complete it completely.

### Is it possible to improve AMFNet?

Our proposed AMFNet improves the performance of SSC by mainly leveraging 2D semantic segmentation results to provide guidance information and applying attention mechanisms to enhance the capability of feature extraction. Here we envision potential improvements in two aspects.

(1) In Table 1 and Table 2, we report the performance of the architecture which uses the ground truth of the 2D semantic segmentation as input to the 3D-guidance branch during the inferring stage (denoted as Ours (Seg-GT)). It can be seen that the ground truth segmentation can improve the performance of the proposed model by 1.8% and 1.2% on the SUNCG-RGBD and NYUv2 datasets respectively, which indicates an accessible way to improve our model is to enhance the performance of the 2D segmentation network.

(2) In the 3D-guidance branch we perform a one-hot encoding to the semantic segmentation results of the visible voxels in a 3D scene, which may lead to inaccurate object constraints for the scenes where the objects belonging to the same category are spatially separated. In these scene samples, a large range of empty voxels between these objects might be encoded into the corresponding category and lead to inaccurate object boundaries. Therefore, another accessible way to improve our model is to use instance segmentation information for one-hot encoding.

### Conclusion

In this paper, we introduce a novel network named AMFNet for the task of semantic scene completion. With the guidance of 2D semantic segmentation and the help of the attention focusing on reliable depth cues in the spatial dimension, AMFNet is capable of improving the completion and segmentation accuracy simultaneously by making full use of the information of the input RGB-D images. The major technique contributions of AMFNet include a two-branch fusion network boosted with 2D semantic segmentation and 3D semantic completion network empowered by residual attention blocks. Experimental results on the SUNCG-RGBD dataset and NYUv2 dataset demonstrate AMFNet achieves state-of-the-art performance. Ablation study and visualization results also show that the two-branch network and the attention-injected network have significant contributions to the proposed method.

## Acknowledgements

This work was supported by the National Natural Science Funds of China (61671267).

## References

- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40:834–848.
- Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision*, 801–818.
- Firman, M.; Mac Aodha, O.; Julier, S.; and Brostow, G. J. 2016. Structured Prediction of Unobserved Voxels from a Single Depth Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5431–5440.
- Geiger, A., and Wang, C. 2015. Joint 3D Object and Layout Inference from a Single RGB-D Image. In *German Conference on Pattern Recognition*, 183–195.
- Graham, B.; Engelcke, M.; and van der Maaten, L. 2018. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9224–9232.
- Guedes, A. B. S.; de Campos, T. E.; and Hilton, A. 2018. Semantic Scene Completion Combining Colour and Depth: Preliminary Experiments. *arXiv preprint arXiv:1802.04735*.
- Guo, Y.-X., and Tong, X. 2018. View-volume Network for Semantic Scene Completion from a Single Depth Image. *arXiv preprint arXiv:1806.05361*.
- Gupta, S.; Arbelaez, P.; and Malik, J. 2013. Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 564–571.
- Gupta, S.; Girshick, R.; Arbeláez, P.; and Malik, J. 2014. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In *Proceedings of the European Conference on Computer Vision*, 345–360.
- Han, X.; Zhang, Z.; Du, D.; Yang, M.; Yu, J.; Pan, P.; Yang, X.; Liu, L.; Xiong, Z.; and Cui, S. 2019. Deep Reinforcement Learning of Volume-guided Progressive View Inpainting for 3D Point Scene Completion from a Single Depth Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 234–243.
- Hays, J., and Efros, A. A. 2007. Scene Completion Using Millions of Photographs. *ACM Transactions on Graphics* 26(3):4.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.
- Laugier, C.; Paromtchik, I. E.; Perrollaz, M.; Yong, M.; Yoder, J.-D.; Tay, C.; Mekhnacha, K.; and Nègre, A. 2011. Probabilistic Analysis of Dynamic Scenes and Collision Risks Assessment to Improve Driving Safety. *IEEE Intelligent Transportation Systems Magazine* 3(4):4–19.
- Li, H.; Xiong, P.; Fan, H.; and Sun, J. 2019a. Dfanet: Deep Feature Aggregation for Real-time Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9522–9531.
- Li, J.; Liu, Y.; Gong, D.; Shi, Q.; Yuan, X.; Zhao, C.; and Reid, I. 2019b. RGBD Based Dimensional Decomposition Residual Network for 3D Semantic Scene Completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7693–7702.
- Lin, D.; Fidler, S.; and Urtasun, R. 2013. Holistic Scene Understanding for 3D Object Detection with RGBD Cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, 1417–1424.
- Liu, S.; Hu, Y.; Zeng, Y.; Tang, Q.; Jin, B.; Han, Y.; and Li, X. 2018. See and Think: Disentangling Semantic Scene Completion. In *Advances in Neural Information Processing Systems*, 263–274.
- Ren, X.; Bo, L.; and Fox, D. 2012. RGB-(D) Scene Labeling: Features and Algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2759–2766.
- Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *Proceedings of the European Conference on Computer Vision*, 746–760.
- Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic Scene Completion from a Single Depth Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1746–1754.
- Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual Attention Network for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.
- Woo, S.; Park, J.; Lee, J.-Y.; and So Kweon, I. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision*, 3–19.
- Xiong, Y.; Liao, R.; Zhao, H.; Hu, R.; Bai, M.; Yumer, E.; and Urtasun, R. 2019. Upsnet: A Unified Panoptic Segmentation Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8818–8826.
- Yang, M.; Yu, K.; Zhang, C.; Li, Z.; and Yang, K. 2018. Denseaspp for Semantic Segmentation in Street Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3684–3692.
- Zhang, J.; Zhao, H.; Yao, A.; Chen, Y.; Zhang, L.; and Liao, H. 2018. Efficient Semantic Scene Completion Network with Spatial Group Convolution. In *Proceedings of the European Conference on Computer Vision*, 733–749.