3D Sketch-aware Semantic Scene Completion via Semi-supervised Structure Prior

Xiaokang Chen[1*] Kwan-Yee Lin[2] Chen Qian[2] Gang Zeng[1†] Hongsheng Li[3]
[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[2]SenseTime Research [3]The Chinese University of Hong Kong

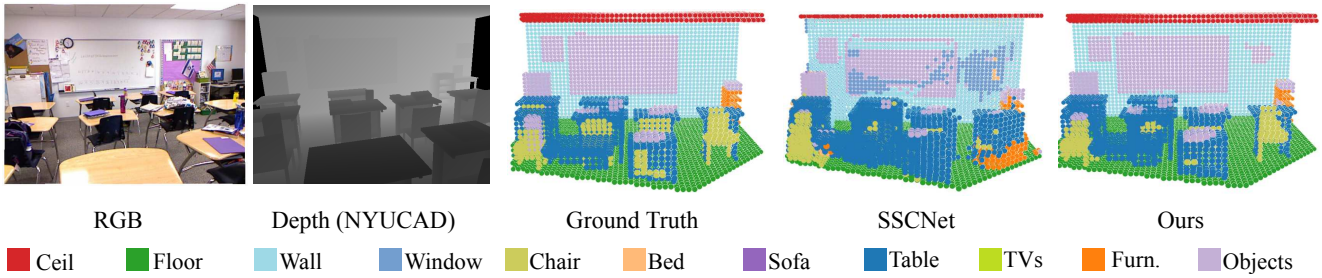

Figure 1. **Visualization of Semantic Scene Completion task.** From left to right: (1) RGB input, (2) depth map, (3) ground truth of semantic scene completion, (4) result of SSCNet [27], (5) result of the proposed method. Our method generates a more reasonable result and obtains a better intra-class consistency and inter-class distinction compared with SSCNet [27], a classic method that models context on implicitly embedded depth feature that learnt from general 3D CNNs.

Abstract

The goal of the Semantic Scene Completion (SSC) task is to simultaneously predict a completed 3D voxel representation of volumetric occupancy and semantic labels of objects in the scene from a single-view observation. Since the computational cost generally increases explosively along with the growth of voxel resolution, most current state-of-the-arts have to tailor their framework into a low-resolution representation with the sacrifice of detail prediction. Thus, voxel resolution becomes one of the crucial difficulties that lead to the performance bottleneck.

In this paper, we propose to devise a new geometry-based strategy to embed depth information with low-resolution voxel representation, which could still be able to encode sufficient geometric information, e.g., room layout, object's sizes and shapes, to infer the invisible areas of the scene with well structure-preserving details. To this end, we first propose a novel 3D sketch-aware feature embedding to explicitly encode geometric information effectively and efficiently. With the 3D sketch in hand, we further devise a simple yet effective semantic scene completion framework that incorporates a light-weight 3D Sketch Hallucination module to guide the inference of occupancy and the semantic labels via a semi-supervised structure prior learning strategy. We demonstrate that our proposed geometric embedding works better than the depth feature learning from habitual SSC frameworks. Our final model surpasses state-of-the-arts consistently on three public benchmarks, which only requires 3D volumes of $60 \times 36 \times 60$ *resolution for both input and output.*

1. Introduction

Semantic Scene Completion (SSC), which provides an alternative to understand the 3D world with both 3D geometry and semantics of the scene from a partial observation, is an emerging topic in computer vision for its wide applicability on many applications, *e.g.*, augmented reality, surveillance and robotics. Due to the high memory and computational cost requirements on inherent voxel representation, most existing methods [27, 9, 41, 7, 14, 16, 40, 4] achieve semantic scene completion through sophisticated 3D context modeling on *implicitly embedded depth feature* that learnt from general 3D CNNs. These methods are either error-prone on classifying fine details of objects or have the difficulties in completing the scene when there exists a large portion of geometry missing, as shown in Figure 1.

Several recent studies [7, 14, 16] present promising re-

* This work was done during an internship at SenseTime Research.
† Gang Zeng is the corresponding author.

sults on this topic by introducing high-resolution RGB images into the process. Though driven by various motivations, these methods could be thought as building *cross-modality feature embedding* with the assumption that the fine detail feature could be compensated from RGB counterpart and computation-efficient property could be guaranteed with 2D operators on RGB source. However, such an approach is highly relied on the effectiveness of cross-modality feature embedding module design and is vulnerable to complex scenes.

In contrast, from the human perception, it is a breeze to complete and recognize 3D scene even from the partial low-resolution observation, due to the prior knowledge on object's geometry properties, *e.g.*, size and shape, of different categories. From this perspective, we hypothesize the feature embedding strategy that explicitly encodes the geometric information could facilitate the network learning the concept of object's structure, and therefore reconstructing and recognizing the scene precisely even from the low-resolution partial observation. To this end, the geometry properties need to be resolution-invariant or at least resolution-insensitive.

Based on this intuition, we present 3D sketch[1]-aware feature embedding, an explicit and compact depth feature embedding schema for the semantic scene completion task. It has been demonstrated in [23] that the similar geometric cue in image space, *i.e.*, 2D boundary, is resolution-insensitive. We show that the 3D world also holds the same conclusion, as indicated in Figure 2.

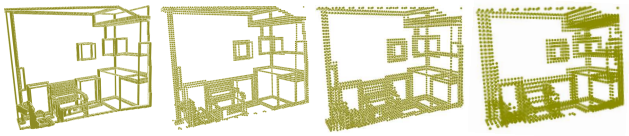

Figure 2. **Visualization of sketches extracted from semantic labels with different resolutions.** From left to right, the sketch begins to lose some details as resolution decreases, while the structure description of the scene is well preserved.

However, 3D sketch extracted from 2D depth image is still a 2D/2.5D observation from a single viewpoint. To fully utilize the strength of this new feature embedding, we further propose a 3D sketch-aware semantic scene completion network, which injects a 3D Sketch Hallucination Module to infer the full 3D sketch from the partial one at first, and then utilize the feature embedded from the hallucinated 3D sketch to guide the reconstruction and recognition. Specifically, since lifting the 2D/2.5D observation to full 3D sketch is intrinsically ambiguous, instead of directly regressing the ground-truth full 3D sketch, we seek a nature prior distribution to sample diverse reasonable 3D sketches. We achieve that by tailoring Conditional Variational Autoencoder (CVAE) [26] into the 3D Sketch Hallucination Module design. We show that such a design could help to generate accurate and realistic results even when there is a large portion of geometry missing from the partial observation.

We summarize our contributions as follows:

- We devise a new geometric embedding from depth information, namely 3D sketch-aware feature embedding, to break the performance bottleneck of the SSC task caused by a low-resolution voxel representation.
- We introduce a simple yet effective semantic scene completion framework that incorporates a novel 3D Sketch Hallucination Module to guide the full 3D sketch inference from partial observation via semi-supervised structure prior property of Conditional Variational Autoencoder (CVAE), and utilizes the feature embedded from the hallucinated 3D sketch to further guide the scene completion and semantic segmentation.
- Our model outperforms state-of-the-arts consistently on three public benchmarks, with only requiring 3D volumes of $60 \times 36 \times 60$ resolution for both input and output.

2. Related Work

2.1. Object Shape Completion

Object shape completion has a long history in geometry processing. We summarize existing methods to two categories: knowledge-based and learning-based.

Knowledge-based methods complete partial input of an object by reasoning geometric cues or matching it with 3D models from an extensive shape database. Some works detect symmetries in meshes or point clouds and use them to fill in missing data, such as [31, 28, 19]. An alternative is to match the partial input with CAD models from a large database [18, 21, 13]. However, it is too expensive to retrieval, and it has poor generalization for new shapes that do not exist in the database.

Learning-based methods are more flexible and effective than knowledge-based ones. They usually infer the invisible area with a deep neural network, which has fast inference speed and better robustness. [2] proposes a 3D-Encoder-Predictor Network, which first encodes the known and unknown space to get a relatively low-resolution prediction, and then correlates this intermediary result with 3D geometry from a shape database. [37] proposes an end-to-end method that directly operates on raw point clouds without any structural assumption about the underlying shape. [29] proposes a weakly-supervised approach that learns a shape

[1] 3D Sketch could be understood as a kind of 3D boundary. To distinguish it with the concept of *edge/boundary* in image space, we refer it as *3D Sketch*.

prior on synthetic data and then conducts maximum likelihood fitting using deep neural networks.

These methods focus on reconstructing 3D shape from the partial input of a single object, which makes it hard for them to extend to partial scenes along with multiple objects estimated in semantic level.

2.2. Semantic Scene Completion

Semantic Scene Completion (SSC) is a fundamental task in 3D scene understanding, which produces a complete 3D voxel representation of volumetric occupancy and semantic labels. SSCNet [27] is the first to combine these two tasks in an end-to-end way. ESSCNet [39] introduces Spatial Group Convolution (SGC) that divides input volume into different groups and conduct 3D sparse convolution on them. VVNet [9] combines 2D and 3D CNN with a differentiable projection layer to efficiently reduce computational cost and enable feature extraction from multi-channel inputs. ForkNet [33] proposes a multi-branch architecture and draws on the idea of generative models to sample new pairs of training data, which alleviates the limited training samples problem on real scenes. CCPNet [41] proposes a self-cascaded context aggregation method to reduce semantic gaps of multi-scale 3D contexts and incorporates local geometric details in a coarse-to-fine manner.

Some works also utilize RGB images as vital complementary to depth. TS3D [7] designs a two-stream approach to leverage semantic and depth information, fused by a vanilla 3DCNN. SATNet [16] disentangles semantic scene completion task by sequentially accomplishing 2D semantic segmentation and 3D semantic scene completion tasks. DDRNet [14] proposes a light-weight Dimensional Decomposition Residual network and fused multi-scale RGB-D features seamlessly.

Above methods could be regraded as encoding depth information implicitly by either single- or cross-modality feature embedding. They map depth information into an inexplicable high-dimensional feature space and then use the feature to predict the result directly. Different from current methods, we propose an explicit geometric embedding strategy from depth information, which predicts 3D sketch first and utilize the feature embedded from it to guide the reconstruction and recognition.

2.3. 2D Boundary Detection

2D Boundary detection is a fundamental challenge in computer vision. There are lots of methods proposed to detect boundaries. Sobel operator [25] and Canny operator [1] are two hand-craft based classics that detect boundaries with gradients of the image. Learning-based works [17, 10, 35] try to employ deep neural networks with supervision. Most of them directly concatenate multi-level features to extract the boundary. Since boundary includes a distinct geometric structure of objects, some other works try to inject boundary detection into other tasks to help boost the performance. [32] combines boundary detection with salient object detection task to encourage better edge-preserving salient object segmentation. [36, 30] introduce boundary detection into semantic segmentation task to obtain more precise semantic segmentation results. [34] achieves robust facial landmark detection by utilizing facial boundary as an intermediate representation to remove the ambiguities. With similar spirits, we introduce a 3D sketch-aware feature embedding to break the performance bottleneck of the SSC task caused by a low-resolution voxel representation.

2.4. Structure Representation Learning

Deep generative models have demonstrated significant performance in structure representation learning. [26] develops a deep conditional generative model to predict structured output using Gaussian latent variables, which can be trained efficiently in the framework of stochastic gradient variational Bayes. [42] proposes an autoencoding formulation to discover landmarks as explicit structural representations in an unsupervised manner. [5] proposes to synthesize images under the guidance of shape representations and conditions on the learned textural information. [22] employs CVAE to stress the issue of the inherent ambiguity in 2D-to-3D lifting in the pose estimation task. Adopting the idea of structure representation learning, we embed the geometric structure of a 3D scene through a CVAE [26] conditioned on the estimated sketch.

3. Methodology

The overall architecture of our network is illustrated in Figure 3. The proposed method consists of multiple stages and each stage adopts an encoder-decoder architecture. Taking a pair of RGB and depth images of a 3D scene as input, the network outputs a dense prediction and each voxel in the view frustum is assigned with a semantic label C_i, where $i \in [0, 1, \cdots, N]$ and N is the number of semantic categories. C_0 stands for empty voxels.

More specifically, we stack two stages and let each stage handle different tasks. The first stage tackles the task of sketch extraction. It embeds the geometric cues contained in the scene and provides the structure prior information (which we call it sketch) for the next stage. Besides, we employ CVAE to guide the predicted sketch. The second stage tackles the task of semantic scene completion (SSC) based on the extracted sketch. Details are introduced below.

3.1. Generation of Ground-truth Sketch

We perform 3D Sobel operator on the semantic label to extract the sketch of the semantic scene. Suppose we have obtained gradients g_x^i, g_y^i, g_z^i at the i-th voxel V_i along x, y, z axes, we first binarize these values to be 0 or 1

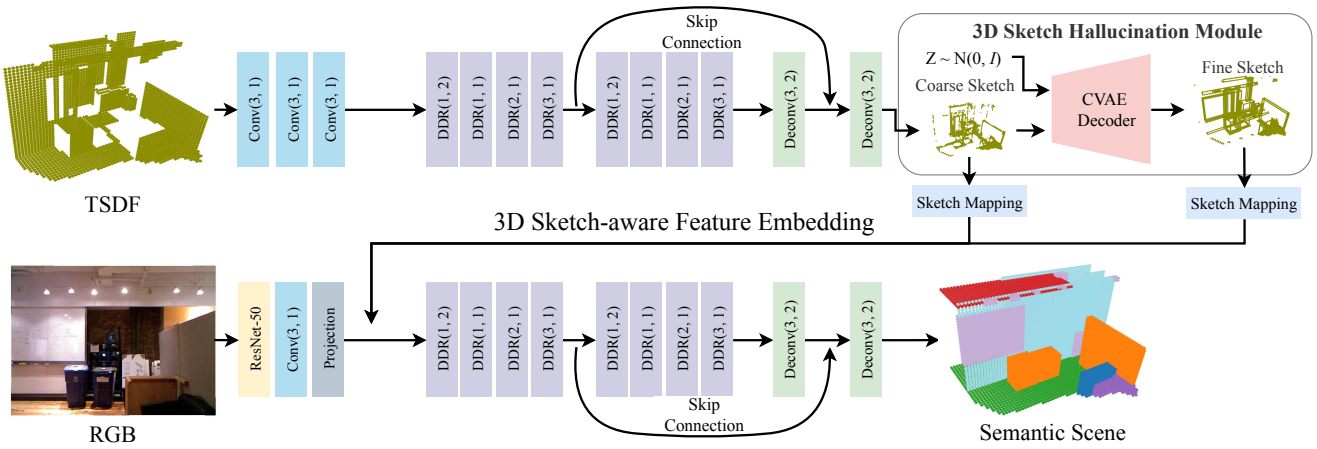

Figure 3. **Overview of our network**. We first generate structure prior information from the TSDF input and use CVAE to refine the prediction. Then the prior information will be passed to the RGB-branch to predict occupancy and object labels for each voxel in the view frustum. The convolution parameters are shown as (kernel size, dilation). The DDR parameters are shown as (dilation, downsample rate). The Deconvolution parameters are shown as (kernel size, upsample rate).

to eliminate the semantic gap. For example, the gap between class 1 and class 2 should be considered equal to the gap between class 1 and class 10 when generating the sketch. Finally, the extracted sketch can be described as a set: $\mathbf{S_{sketch}} = \{V_i : g_x^i + g_y^i + g_z^i > 1\}$. To distinguish generated geometric representation with generally 2D edge/boundary, we refer it as *3D Sketch*.

3.2. Sketch Prediction Stage

This stage takes a single-view depth map as input and encodes it as a 3D volume. We follow [27] to rotate the scene to align with gravity and room orientation based on Manhattan assumption. We adopt Truncated Signed Distance Function (TSDF) to encode the 3D space, where every voxel stores the distance value d to its closest surface and the sign of the value indicates whether the voxel is in free space or occluded space. The encoder volume has a grid size of 0.02 m and a truncation value of 0.24 m, resulting in a $240 \times 144 \times 240$ volume. For the saving of computational cost, [27] downsamples the ground truth by a rate of 4, and we use the same setting. Following SATNet [16], we also downsample the input volume by a rate of 4 and use $60 \times 36 \times 60$ resolution as input.

Previous works [20, 38, 36] demonstrate that contextual information is important for 2D semantic segmentation. Due to the sparseness and the high computational cost of 3D voxels, it is hard to obtain the context of the scene. To learn rich contextual information, we should make sure that our network has a large enough receptive field without significantly increasing the computational cost. To this end, [14] proposed Dimensional Decomposition Residual (DDR) block which is computation-efficient compared with basic 3D residual block. We adopt DDR block as our basic unit and stack them layer by layer with different dilation rates to maintain big receptive fields. As shown in Figure 3, We first employ several convolutions to encode the TSDF volume into high dimensional features. Then we aggregate the contextual information of the input feature by several DDR blocks and downsample it by a rate of 4 to reduce computational cost. Finally, we employ two deconvolution layers to upsample the feature volume and obtain the dense predicted sketch, which we denote as $\hat{G}_{raw}$. Following [27], we add a skip connection between two layers for better gradient propagation, which is illustrated in Figure 3.

Due to the input of semantic scene completion task is not a complete scene, we assume that a more precise and complete sketch will bring more information increments to the subsequent stage. To some extent, it may make up for the inadequacy of incomplete input. Thus we design a 3D Sketch Hallucination Module to handle this issue.

3.3. 3D Sketch Hallucination Module

Lifting 2D/2.5D observation to full 3D sketch is intrinsically ambiguous, we thus seek a nature prior distribution to sample diverse reasonable 3D sketches instead of directly regressing the ground truth. Thus, we employ CVAE to further process the original predicted sketch by sampling an accurate and diverse sketch set $S = \{\hat{G}^k_{refined} : k \in 1, 2, ..., K\}$ conditioned on the estimated $\hat{G}_{raw}$.

The proposed 3D Sketch Hallucination Module (as shown in Figure 4) consists of a standard encoder-decoder structure. The encoder which we denote as $\mathcal{E}(G_{gt}, \hat{G}_{raw})$, performs some convolution operations on the input ground-truth sketch and a condition $\hat{G}_{raw}$ to output the mean and diagonal covariance for the posterior $q(\hat{z}|G_{gt}, \hat{G}_{raw})$. Then the decoder which we denote as $\mathcal{D}(\hat{z}, \hat{G}_{raw})$ will reconstruct the sketch by taking a latent $\hat{z}$ sampled from the pos-

terior $q(\hat{z}|G_{gt}, \hat{G}_{raw})$ and the condition $\hat{G}_{raw}$ as input.

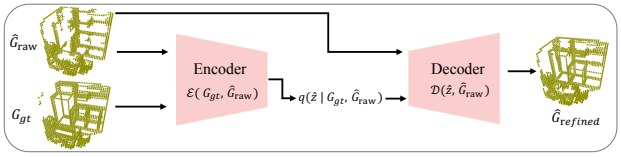

Figure 4. **Architecture of the proposed Sketch Hallucination Module**. During training time, the original estimated sketch and the ground-truth sketch are fed into the encoder to generate mean and diagonal covariance for the posterior q. Then the decoder will reconstruct the ground-truth sketch with a latent sampled from q and the original estimated sketch as input.

During training, we optimize the proposed module though minimizing the following objective function,

$$\begin{aligned}\mathcal{L}_{\text{CVAE}} = \lambda_1 KL(q(\hat{z}|G_{gt}, \hat{G}_{raw}) \,||\, p(z|\hat{G}_{raw})) \\ + \lambda_2 \mathbb{E}_{z\sim q(\hat{z}|G_{gt}, \hat{G}_{raw})} \boldsymbol{\epsilon}(G_{gt}, \mathcal{D}(\hat{z}, \hat{G}_{raw})),\end{aligned} \tag{1}$$

where $\boldsymbol{\epsilon}$ is a cross-entropy loss and $KL(x||y)$ is the Kullback-Leibler divergence loss. We use λ_i as hyper-parameter to weight these two loss items. $\mathbb{E}$ is the expectation which is taken over K samples. The $p(z|\hat{G}_{raw})$ is the prior distribution. To ensure gradients can be backpropagated through the latent code, the KL divergence is required to be computed in a closed form. Thus, the latent space of CVAE is typically restricted to be a distribution over $\mathcal{N}(0, I)$. We follow this setting in our framework. Specifically, it draws a Gaussian prior assumption over the coarse-step geometry representation to fine-step geometry representation in our framework. Sketch is a simple yet compact geometry representation which suits the assumption. Since the encoder will not be used during inference, the current objective will introduce inconsistency between training and inference. To address this issue, we follow [26, 22] to set the encoder the same as prior network $p(z) \sim \mathcal{N}(0, I)$, namely *Gaussian Stochastic Neural Network* (GSNN) and the reparameterization trick of CVAE can be used to train GSNN. We combine $\mathcal{L}_{\text{GSNN}}$ and $\mathcal{L}_{\text{CVAE}}$ with α as weight term to obtain final objective for our refine network,

$$\mathcal{L}_{\text{GSNN}} = \mathbb{E}_{z\sim\mathcal{N}(0,I)} \boldsymbol{\epsilon}(G_{gt}, \mathcal{D}(z, \hat{G}_{raw})), \tag{2}$$

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{CVAE}} + \alpha \mathcal{L}_{\text{GSNN}}, \tag{3}$$

Durning inference, we randomly sample z from $\mathcal{N}(0, I)$ for K times and obtain K different $\mathcal{D}(z, \hat{G}_{raw})$, which are denoted as $S = \{\hat{G}^k_{refined} : k \in 1, 2, ..., K\}$. We average them and obtain the refined sketch $\hat{G}_{refined}$.

3.4. Semantic Scene Completion Stage

In this stage, we will take a single RGB image and the pre-computed sketches from the former stage as input to densely predict the semantic scene labels. We divide this stage into three parts: 2D feature learning, 2D-3D projection and 3D feature learning. The input RGB image is firstly fed into a ResNet-50 [11] to extract local and global textural features. For achieving stable training, we utilize the parameters pre-trained on ImageNet [3] and freeze the weight of them. Due to the output tensor of ResNet-50 has too many channels, which will bring too much computational cost for 3D learning part, we adopt a convolution layer followed by a Batch Normalization [12] and Rectified Linear Unit (ReLU) to reduce its dimensions.

Then the computed 2D semantic feature map will be projected into 3D space according to the depth map and the corresponding camera parameters. Given the depth image I_{depth}, the intrinsic camera matrix $K_{camera} \in \mathbb{R}^{3\times3}$, and the extrinsic camera matrix $E_{camera} \in \mathbb{R}^{3\times4}$, each pixel $p_{u,v}$ in the 2D feature map can be projected to an individual 3D point $p_{x,y,z}$. Because the resolution of the 3D volume is lower than the 2D feature map, multiple points may be divided into the same voxel in the process of voxelization. For those voxels, we only keep one feature vector in a certain voxel by max-pooling. After this step, the semantic feature vector for each pixel is assigned to its corresponding voxel via the mapping $\mathbb{M}$. Since many areas are not visible, zero vectors are assigned to the occluded areas and empty foreground in the scene.

Given the projected 3D feature map $\mathbf{F}_{proj} \in \mathbb{R}^{C\times H\times W\times L}$, where C is the number of channels and H, W, L are size of the feature map. We now use the prior information $\hat{G}_{raw}$ and $\hat{G}_{refined}$ as guidance. We define two sketch mappings: $\mathcal{F}_{raw} : \hat{G}_{raw} \rightarrow \mathbf{F}_{raw} \in \mathbb{R}^{C\times H\times W\times L}$ and $\mathcal{F}_{refined} : \hat{G}_{refined} \rightarrow \mathbf{F}_{refined} \in \mathbb{R}^{C\times H\times W\times L}$ to map these prior information to the same feature space with $\mathbf{F}_{proj}$. After these two mapping operations, both $\mathbf{F}_{raw}$ and $\mathbf{F}_{refined}$ have the same resolution and dimension with $\mathbf{F}_{proj}$. Thus we introduce the prior information by an element-wise addition operation on $\mathbf{F}_{proj}$, $\mathbf{F}_{raw}$ and $\mathbf{F}_{refined}$. In pratice, these two mapping functions are implemented by 3×3 convolution layers. In the following, the new feature map will be fed into a 3D CNN, whose architecture is the same with that of sketch-branch, and we obtain the final semantic scene completion predictions.

3.5. Loss Function

During training, the dataset is organized as a set $\{(X_{\text{TSDF}}, X_{\text{RGB}}, G_{gt}, S_{gt})\}$, where G_{gt} represents the ground-truth sketch and S_{gt} represents the ground-truth semantic labels. We optimize the entire architecture by the following formulas:

$$\mathcal{L}_{\text{loss}} = \mathcal{L}_{\text{semantic}} + \mathcal{L}_{\text{hybrid}} + \mathcal{L}_{\text{sketch}}, \tag{4}$$

$$\mathcal{L}_{\text{semantic}} = \boldsymbol{\epsilon}(S_{gt}, \mathcal{D}_s(\mathcal{E}_s(X_{\text{RGB}}))), \tag{5}$$

$$\mathcal{L}_{\text{sketch}} = \boldsymbol{\epsilon}(G_{gt}, \mathcal{D}_g(\mathcal{E}_g(X_{\text{TSDF}})), \tag{6}$$

where $\mathcal{D}_g, \mathcal{E}_g$ are the encoder and the decoder of the sketch stage, $\mathcal{D}_s, \mathcal{E}_s$ are the encoder and the decoder of the semantic stage, $\mathcal{L}_{\text{hybrid}}$ is defined in Eq. (3), and ϵ denotes the cross-entropy loss.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate the proposed method on three datasets: NYU Depth V2 [24] (which is denoted as NYU in the following), NYUCAD [6] and SUNCG [27]. We will introduce these three datasets in detail in the supplementary material. We follow SSCNet [27] and use precision, recall and voxel-level intersection over union (IoU) as evaluation metrics. Following [27], two tasks are considered: semantic scene completion (SSC) and scene completion (SC). For the task of SSC, we evaluate the IoU of each object class on both observed and occluded voxels in the view frustum. For the task of SC, we treat all voxels as binary predictions, $i.e.$, empty or non-empty. We evaluate the binary IoU on occluded voxels in the view frustum.

4.2. Implementation Details

Training Details. We use PyTorch framwork to implement our experiments with 2 GeForce GTX 1080 Ti GPUs. We adopt mini-batch SGD with momentum to train our model with batch size 4, momentum 0.9 and weight decay 0.0005. We employ a poly learning rate policy where the initial learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{0.9}$. For both NYU and NYUCAD, we train our network for 250 epochs with initial learning rate 0.1. For SUNCG, we train our network for 8 epochs with initial learning rate 0.01. The expection in Eq. (1) is estimated using $K = 4$ samples. λ_1, λ_2 and α in Eq. (1) and Eq. (3) are set to 2, 1 and 1.5 respectively.

Drop Rate(%)	SC-IoU(%)	SSC-mIoU(%)
0	**94.2**	**65.0**
20	93.7	63.6
40	93.2	62.3
60	92.0	59.9
80	89.9	57.1

Table 1. **Oracle Ablation**. *(Oracle) Drop Rate* means we randomly drop the ground-truth sketch in a certain proportion. We perform this ablation study on NYUCAD dataset.

Oracle Ablation. To obtain the theoretical upper limit of the proposed method, we replace the output of the first stage with the ground-truth 3D sketch to supply the structure prior. Results are shown in Table 1. *Drop Rate* means we randomly discard some voxels in the ground-truth 3D sketch by some ratio. We observe that with the whole 3D sketch as structure prior, our network could infer most of the invisible areas and obtain 94.2% SC IoU. As the drop rate increases to 80%, the performance has not dropped a lot and is still higher than the best performance of the proposed method, which verifies the validity of accurate structure prior.

4.3. Comparisons with State-of-the-art Methods

We further compare the proposed method with state-of-the-art methods. Table 3 shows the performances by state-of-the-art methods on NYU dataset. We observe that the proposed method outperforms all existing methods by a large margin, more specifically, we gain an increase of 7.8% SC IoU and 2.6% SSC mIoU compared to CCPNet [41]. We argue that this improvement is caused by the novel two-stage architecture which makes the full use of the structure prior. The provided structure prior can accurately infer invisible areas of the scene with well structure-preserving details.

We also conduct experiments on NYUCAD dataset to validate the generalization of the proposed method. Table 4 presents the quantitative results on NYUCAD dataset. Our proposed method maintains the performance advantage and outperforms CCPNet [41] by 1.8% SC IoU and 2.0% SSC mIoU. Note that although some works [41, 33, 7] use larger input resolution than ours, the proposed method still outperforms them with a low-resolution input of $60 \times 36 \times 60$.

Experiments on SUNCG dataset and the visualization of the SSC results compared with SSCNet [27] on NYUCAD dataset are put in the supplementary material.

4.4. Ablation Study

To evaluate the effectiveness of the pivotal components of our method, we perform extensive ablation studies using the same hyperparameters. Details are illustrated below.

#Stage	Structure Prior	CVAE	SC-IoU(%)	SSC-mIoU(%)
1	✗	✗	79.3	48.7
2	✗	✗	81.1	50.6
2	✓	✗	83.6	53.9
2	✓	✓	**84.2**	**55.2**

Table 2. **Ablation studies on different modules**. We perform this ablation study on NYUCAD dataset.

Different Modules in the Framework. We first conduct ablation studies on different modules in the proposed method. Results are shown in Table 2. From Row 1 and Row 2, we find that just adopting a dual-path structure could boost the performance, as more parameters are introduced. In the third row, with the introduction of structure prior, our network could infer the invisible areas of the scene with well structure-preserving details, which brings great improvements. Finally, with the proposed 3D Sketch Hallucination Module, we further boost the performance and achieve 84.2% SC IoU and 55.2% SSC mIoU, which are both new state-of-the-art performance on NYUCAD.

Different Representations of Structure Prior. We also perform ablation studies on different representations of

			scene completion			semantic scene completion											
Methods	Resolution	Trained on	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
Lin *et al.* [15]	$(240, 60)$	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 *et al.* [8]	$(240, 60)$	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
SSCNet [27]	$(240, 60)$	NYU	57.0	**94.5**	55.1	15.1	94.7	24.4	0.0	12.6	32.1	35.0	13.0	7.8	27.1	10.1	24.7
ESSCNet [39]	$(240, 60)$	NYU	71.9	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 [14]*	$(240, 60)$	NYU	71.5	80.8	61.0	21.1	92.2	33.5	6.8	14.8	48.3	42.3	13.2	13.9	35.3	13.2	30.4
VVNetR-120 [9]	$(120, 60)$	NYU+SUNCG	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
TS3D [7]*	$(240, 60)$	NYU	-	-	60.0	9.7	93.4	25.5	21.0	17.4	55.9	49.2	17.0	27.5	39.4	19.3	34.1
SATNet-TNetFuse [16]*	$(60, 60)$	NYU+SUNCG	67.3	85.8	60.6	17.3	92.1	28.0	16.6	19.3	57.5	53.8	17.2	18.5	38.4	18.9	34.4
ForkNet [33]	$(80, 80)$	NYU	-	-	63.4	36.2	93.8	29.2	18.9	17.7	**61.6**	52.9	23.3	19.5	**45.4**	20.0	37.1
CCPNet [41]	$(240, 240)$	NYU	74.2	90.8	63.5	23.5	**96.3**	35.7	20.2	25.8	61.4	**56.1**	18.1	**28.1**	37.8	20.1	38.5
Ours*	$(60, 60)$	NYU	**85.0**	81.6	**71.3**	**43.1**	93.6	**40.5**	**24.3**	**30.0**	57.1	49.3	**29.2**	14.3	42.5	**28.6**	**41.1**

Table 3. **Results on NYU dataset**. Bold numbers represent the best scores. *Resolution(a, b)* means the input resolution is $(a \times 0.6a \times a)$ and the output resolution is $(b \times 0.6b \times b)$. '*' are RGB-D based methods.

			scene completion			semantic scene completion											
Methods	Resolution	Trained on	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
Zheng *et al.* [43]	$(240, 60)$	NYUCAD	60.1	46.7	34.6	-	-	-	-	-	-	-	-	-	-	-	-
Firman *et al.* [6]	$(240, 60)$	NYUCAD	66.5	69.7	50.8	-	-	-	-	-	-	-	-	-	-	-	-
SSCNet [27]	$(240, 60)$	NYUCAD+SUNCG	75.4	**96.3**	73.2	32.5	92.6	40.2	8.9	33.9	57.0	59.5	28.3	8.1	44.8	25.1	40.0
VVNetR-120 [9]	$(120, 60)$	NYUCAD+SUNCG	86.4	92.0	80.3	-	-	-	-	-	-	-	-	-	-	-	-
DDRNet [14]*	$(240, 60)$	NYUCAD	88.7	88.5	79.4	54.1	91.5	56.4	14.9	37.0	55.7	51.0	28.8	9.2	44.1	27.8	42.8
TS3D [7]*	$(240, 60)$	NYUCAD	-	-	76.1	25.9	93.8	48.9	33.4	31.2	66.1	56.4	31.6	**38.5**	51.4	30.8	46.2
CCPNet [41]	$(240, 240)$	NYUCAD	**91.3**	92.6	82.4	56.2	**94.6**	58.7	**35.1**	44.8	68.6	65.3	37.6	35.5	53.1	35.2	53.2
Ours*	$(60, 60)$	NYUCAD	90.6	92.2	**84.2**	**59.7**	94.3	**64.3**	32.6	**51.7**	**72.0**	**68.7**	**45.9**	19.0	**60.5**	**38.5**	**55.2**

Table 4. **Results on NYUCAD dataset**. Bold numbers represent the best scores. *Resolution(a, b)* means the input resolution is $(a \times 0.6a \times a)$ and the output resolution is $(b \times 0.6b \times b)$. '*' are RGB-D based methods.

Input	Shape	Semantic Labels	Sketch	SC-IoU(%)	SSC-mIoU(%)
TSDF+RGB	✓			83.1	52.5
TSDF+RGB		✓		82.6	53.2
TSDF+RGB			✓	**84.2**	**55.2**

Table 5. **Ablation studies on different representations of structure prior**. We perform this ablation study on NYUCAD dataset.

Supervision	Embedding	SC-IoU(%)	SSC-mIoU(%)
None	Implicit	81.1	50.6
Shape	Implicit	83.1	51.8
	Explicit	**83.1**	**52.5**
Semantic	Implicit	82.3	52.1
	Explicit	**82.6**	**53.2**
Sketch	Implicit	83.5	54.4
	Explicit	**84.2**	**55.2**

Table 6. **Ablation studies on different types of embeddings**. We perform this ablation study on NYUCAD dataset.

structure prior. We list three different representations of the prior here: shape, semantic labels and sketch. Shape is the binary description of the scene and we generate the ground-truth shape by binarizing the semantic labels. Semantic labels and sketch have been introduced in the above sections. From Table 5, we observe that sketch is the best representation for modelling structure prior as it could infer the invisible regions with well structure-preserving details.

Different Types of Embeddings. In this part, we conduct ablation studies on different types of embeddings. Results are shown in Table 6. 'Implicit' represents taking the output of the last deconvolution layer in the first stage as the geometric embedding and feed it to the second stage as prior information. 'Explicit' represents we abstract a concrete structure based on the implicit embedding and use it a structure prior. We observe that even using implicit embedding, adding any reasonable supervision on it could boost the performance, such as semantics, shape and sketch. When we convert to explicit embedding, a better structure prior is obtained and the performance shows another boost. Note that the explicit embedding supervised by sketch outperforms its baseline using implicit embedding with no supervision by 3.1% SC IoU and 4.6% SSC mIoU, which demonstrates the effectiveness of the proposed sketch structure prior and the explicit embedding method.

Input for Stage1	Input for Stage2	SC-IoU(%)	SSC-mIoU(%)
RGB	RGB	68.0	40.0
RGB	TSDF	71.2	40.2
TSDF	TSDF	**71.5**	37.2
TSDF	RGB	71.3	**41.1**

Table 7. **Ablation studies on different modal input**. We perform this ablation study on NYU dataset.

Different Modal Input. We adopt data from different modalities as input, more specifically, TSDF for the first stage and RGB for the second stage. We claim that TSDF embeds rich geometric information and is suitable for the sketch prediction task, while RGB is rich in semantic information and is suitable for semantic label prediction task. Results are shown in Table 7. From Row 1 and Row 4, we observe that TSDF generates better structure prior than RGB, resulting in a gain of 3.3% SC IoU. From Row 3 and Row 4, we observe that RGB generates more precise se-

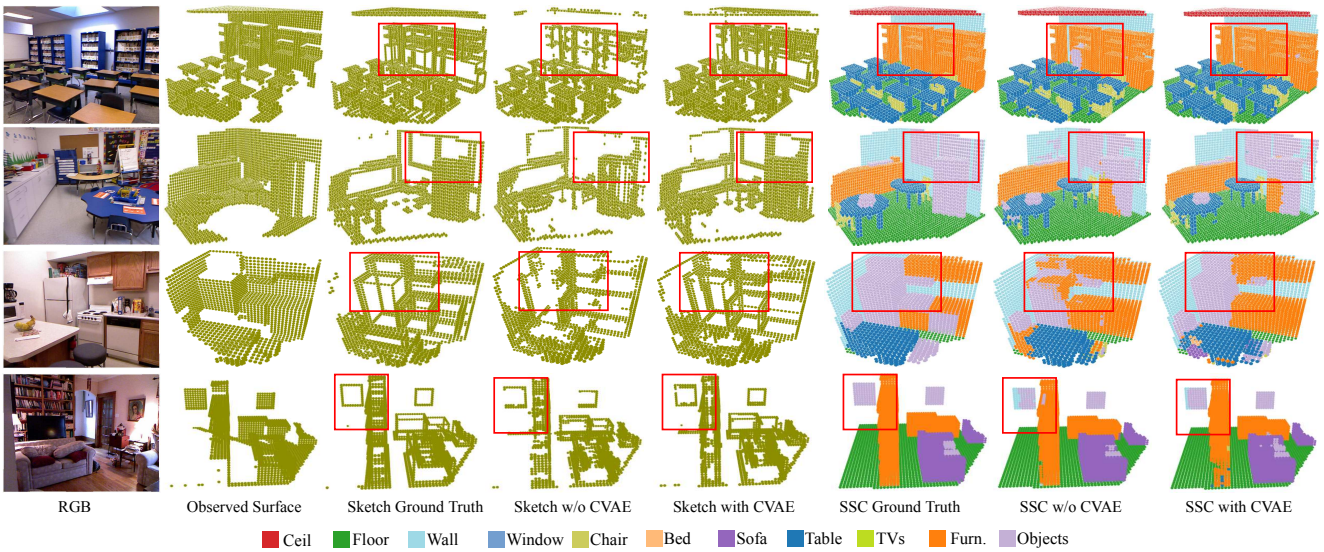

Figure 5. **Visualization of the sketch on NYUCAD dataset.** With the proposed 3D Sketch Hallucination Module, which leverages CVAE to guide the inference of invisible areas, the sketch obtains a sharper boundary and is completer, resulting in better semantic predictions.

Dateset	Resolution	SC-IoU(%)	SSC-mIoU(%)
NYU	(60, 60)	71.3	41.1
NYU	(80, 60)	71.4	**41.2**
NYU	(80, 80)	**76.5**	40.0
NYUCAD	(60, 60)	84.2	55.2
NYUCAD	(80, 60)	84.1	**55.9**
NYUCAD	(80, 80)	**86.0**	54.9

Table 8. **Ablation studies on input/output resolutions**. We perform this ablation study on NYU and NYUCAD dataset both. *Resolution(a, b)* means the input resolution is $(a \times 0.6a \times a)$ and the output resolution is $(b \times 0.6b \times b)$.

mantic labels based on the same structure prior provided by TSDF, resulting in a gain of 3.9% SSC mIoU. From Row 1, Row 2 and Row 3, we observe that the introduction of other modalities would result in corresponding gains on the basis of single-mode data.

Different Input/Output Resolutions. In this part, we conduct ablation studies to verify the impacts of different input/output resolutions on the performance. Results are shown in Table 8. We observe that increasing input size would not make the performance worse. If we increase both the input and output resolutions, SC IoU increases substantially, while SSC mIoU only declines slightly. Hence we conclude that increasing resolution of either input or output is beneficial to semantic scene completion task.

4.5. Qualitative Results of 3D Sketch

We visualize the predicted 3D sketch with/without CVAE in Figure 5. We can observe that the sketch is more complete and precise with the proposed 3D Sketch Hallucination Module. Under the constraints of a more complete sketch, the semantic result shows great consistency in regions with the same semantic labels and has a sharper boundary. For example, in the first row, some regions in the bookcase are mislabeled as $objects$ without CVAE, and those regions in the corresponding sketch are missing. In the second row, the sketch without CVAE fails to extract the outline of the object on the wall, leading to uncertainty of the semantic boundary. In the third row, the missing boundary in the sketch without CVAE brings confusing semantics. In the last row, the sketch of the photo frame is incomplete without CVAE, resulting in more areas to be mislabeled as $wall$.

5. Conclusion

In this paper, we propose a novel 3D sketch-aware feature embedding scheme which explicitly embeds geometric information with structure-preserving details. Based on this, we further propose a semantic scene completion framework that incorporates a novel 3D Sketch Hallucination Module to guide full 3D sketch inference from partial observation via structure prior. Experiments show the effectiveness and efficiency of the proposed method, and state-of-the-art performances on three public benchmarks are achieved.

Acknowledgments: This work is supported by the National Key Research and Development Program of China (2017YFB1002601, 2016QY02D0304), National Natural Science Foundation of China (61375022, 61403005, 61632003), Beijing Advanced Innovation Center for Intelligent Robots and Systems (2018IRS11), and PEK-SenseTime Joint Laboratory of Machine Vision.

References

[1] John Canny. A computational approach to edge detection. *PAMI*, (6):679–698, 1986.

[2] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *CVPR*, pages 5868–5877, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[4] Aloisio Dourado, Teofilo Emidio de Campos, Hansung Kim, and Adrian Hilton. Edgenet: Semantic scene completion from rgb-d images. *arXiv preprint arXiv:1908.02893*, 2019.

[5] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, pages 8857–8866, 2018.

[6] Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J Brostow. Structured prediction of unobserved voxels from a single depth image. In *CVPR*, pages 5431–5440, 2016.

[7] Martin Garbade, Johann Sawatzky, Alexander Richard, and Juergen Gall. Two stream 3d semantic scene completion. 2019.

[8] Andreas Geiger and Chaohui Wang. Joint 3d object and layout inference from a single rgb-d image. In *GCPR*, pages 183–195, 2015.

[9] Yu-Xiao Guo and Xin Tong. View-volume network for semantic scene completion from a single depth image. In *IJCAI*, pages –, 2018.

[10] Jianzhong He, Shiliang Zhang, Ming Yang, Yanhu Shan, and Tiejun Huang. Bi-directional cascade network for perceptual edge detection. In *CVPR*, pages 3828–3837, 2019.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[13] Young Min Kim, Niloy J Mitra, Dong-Ming Yan, and Leonidas Guibas. Acquiring 3d indoor environments with variability and repetition. *TOG*, 31(6):138, 2012.

[14] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *CVPR*, pages –, 2019.

[15] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, pages 1417–1424, 2013.

[16] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. In *NIPS*, pages 261–272, 2018.

[17] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer convolutional features for edge detection. In *CVPR*, pages 3000–3009, 2017.

[18] Liangliang Nan, Ke Xie, and Andrei Sharf. A search-classify approach for cluttered indoor scene understanding. *TOG*, 31(6):137, 2012.

[19] Mark Pauly, Niloy J Mitra, Johannes Wallner, Helmut Pottmann, and Leonidas J Guibas. Discovering structural regularity in 3d geometry. In *TOG*, volume 27, page 43. ACM, 2008.

[20] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *CVPR*, pages 4353–4361, 2017.

[21] Tianjia Shao, Weiwei Xu, Kun Zhou, Jingdong Wang, Dongping Li, and Baining Guo. An interactive approach to semantic modeling of indoor scenes with an rgbd camera. *TOG*, 31(6):136, 2012.

[22] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *CVPR*, pages 2325–2334, 2019.

[23] Yukai Shi, Keze Wang, Chongyu Chen, Li Xu, and Liang Lin. Structure-preserving image super-resolution via contextualized multitask learning. *IEEE transactions on multimedia*, 19(12):2804–2815, 2017.

[24] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[25] I. Sobel and G. Feldman. A computational approach to edge detection, 1968.

[26] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, pages 3483–3491, 2015.

[27] 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 1746–1754, 2017.

[28] Pablo Speciale, Martin R Oswald, Andrea Cohen, and Marc Pollefeys. A symmetry prior for convex variational 3d reconstruction. In *ECCV*, pages 313–328. Springer, 2016.

[29] David Stutz and Andreas Geiger. Learning 3d shape completion from laser scan data with weak supervision. In *CVPR*, pages 1955–1964, 2018.

[30] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *ICCV*, pages 5229–5238, 2019.

[31] Sebastian Thrun and Ben Wegbreit. Shape from symmetry. In *ICCV*, volume 2, pages 1824–1831. IEEE, 2005.

[32] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *CVPR*, pages 1448–1457, 2019.

[33] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *ICCV*, pages 8608–8617, 2019.

[34] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018.

[35] Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming-Hsuan Yang. Object contour detection with a fully convolutional encoder-decoder network. In *CVPR*, pages 193–202, 2016.

[36] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, pages 1857–1866, 2018.

[37] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *3DV*, pages 728–737. IEEE, 2018.

[38] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, pages 7151–7160, 2018.

[39] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *ECCV*, pages 733–749, 2018.

[40] Liang Zhang, Le Wang, Xiangdong Zhang, Peiyi Shen, Mohammed Bennamoun, Guangming Zhu, Syed Afaq Ali Shah, and Juan Song. Semantic scene completion with dense crf from a single depth image. *Neurocomputing*, 318:182–195, 2018.

[41] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *ICCV*, pages 7801–7810, 2019.

[42] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, pages 2694–2703, 2018.

[43] 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.