NDC-Scene: Boost Monocular 3D Semantic Scene Completion in Normalized Device Coordinates Space

Jiawei Yao[1*]   Chuming Li[2,4*]   Keqiang Sun[3*]   Yingjie Cai[3]   Hao Li[3]
Wanli Ouyang[4†] Hongsheng Li[3,4,5†]
[1] University of Washington [2] The University of Sydney
[3] CUHK-SenseTime Joint Laboratory [4] Shanghai AI Laboratory [5] CPII under InnoHK
jwyao@uw.edu, chli3951@uni.sydney.edu.au, wanli.ouyang@sydney.edu.au,
{kqsun@link, caiyingjie@link, haoli@link, hsli@ee}.cuhk.edu.hk

Abstract

Monocular 3D Semantic Scene Completion (SSC) has garnered significant attention in recent years due to its potential to predict complex semantics and geometry shapes from a single image, requiring no 3D inputs. In this paper, we identify several critical issues in current state-of-the-art methods, including the Feature Ambiguity of projected 2D features in the ray to the 3D space, the Pose Ambiguity of the 3D convolution, and the Computation Imbalance in the 3D convolution across different depth levels. To address these problems, we devise a novel Normalized Device Coordinates scene completion network (NDC-Scene) that directly extends the 2D feature map to a Normalized Device Coordinates (NDC) space, rather than to the world space directly, through progressive restoration of the dimension of depth with deconvolution operations. Experiment results demonstrate that transferring the majority of computation from the target 3D space to the proposed normalized device coordinates space benefits monocular SSC tasks. Additionally, we design a Depth-Adaptive Dual Decoder to simultaneously upsample and fuse the 2D and 3D feature maps, further improving overall performance. Our extensive experiments confirm that the proposed method consistently outperforms state-of-the-art methods on both outdoor SemanticKITTI and indoor NYUv2 datasets. Our code are available at https://github.com/Jiawei-Yao0812/NDCScene.

1. Introduction

Semantic Scene Completion (SSC) [38] is a crucial task in 3D scene understanding [43] [20], with wide applications like virtual reality, embodied AI, and autonomous driving,

---

[*]These authors contribute equally to this work.
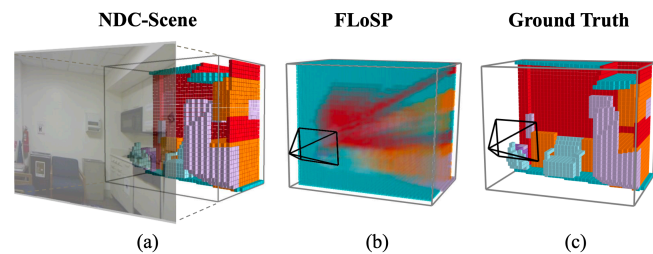[†]Corresponding authors.

Figure 1: **Feature ambiguity**. We compare (a) the feature maps generated by the proposed dual decoder in the normalized device coordinates space, and (b) the feature maps projected through FLoSP [7], with reference to the ground truth demonstrated in (c). In (b), the multi-scale 2D features are projected along their line of sight, which introduces ambiguity in both feature size and feature depth. Conversely, in the normalized device coordinates space, the semantics and occupancy are implicitly restored, as exemplified in (a).

etc. Despite the growing body of research on this topic, the majority of existing SSC solutions [35] [6] [8] depend on input RGB image and corresponding 3D inputs such as depth image, truncated signed distance function (TSDF)[1], *etc*., to forecast volumetric occupancy and the corresponding semantic labels. However, the reliance on these 3D data often entails the use of specialized and costly depth sensors, and thus limits the further application of the Semantic Scene Completion algorithms. Recently, there is growing interest in monocular 3D semantic scene completion [7], which aims to reconstruct a volumetric 3D scene from a single RGB image, thus eliminating the requirement of the additional 3D inputs.

In the pioneering method Monoscene [7], the 2D features are lifted to the 3D space by inverting the perspective projection, where the same 2D features are propagated to

---

[1]TSDF is a representation to encode depth volume, where each voxel stores the distance value to its closest surface and the sign indicates whether the voxel is in visible spaces.
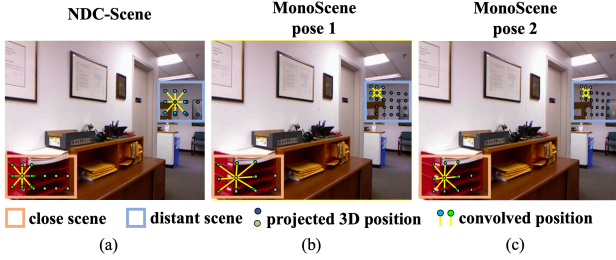
Figure 2: **Pose ambiguity and Imbalanced computation**. We illustrate the projected 2D positions from the 3D convolutions in (a) the proposed normalized device coordinates space and (bc) the target space. We chose two 3D convolution layer with the same stride and resolution, respectively in the two spaces. In (a), the projected positions uniformly distributes on the 2D pixels while in (bc) the positions have an imbalanced allocation between the close and far scenes. Moreover, the convolution scope is not consistent among different choices of camera poses, especially manifesting in the different convolution offsets, as exemplified in (bc).

different depths along the cast camera rays. As depicted in Fig. 1 (b), all the voxels on the projected position of the line sharing the same 2D feature at that specific position. This approach broadcast the 2D feature in the 3D space, making it possible to employ a 3D UNet to predict the completed semantic scene volume.

However, we notice some ambiguities in the prior works, which can be summarized as Feature-Size Ambiguity (FSA), Feature-Depth Ambiguity (FDA) and Pose Ambiguity (PA). The aforementioned projection gives rise to the FSA and FDA. With regard to FSA, the utilization of perspective projection results in the spread of 2D image features across a larger space as the depth increases. As shown in Fig. 1 (b), this leads to variations in feature density across different depths. These 3D features with inconsistent density pose a challenge for the convolution kernel to discern the effective patterns. As for FDA, as shown in Fig. 1 (b), each 2D feature pixel corresponds to a specific position and category, hence propagating 2D feature pixels to all the voxels along the ray makes the depth and category indistinguishable in the constructed 3D features.

The Pose Ambiguity (PA) lies in the lack of camera extrinsic parameters. As illustrated in Fig. 2 (b) and (c), given a certain input image, by supposing different camera extrinsic parameters, the relative positions between the convolved positions of a 3D convolution and the convolution center, when projected on the 2D feature map, should also transform accordingly. In other words, the 3D convolution should be conditioned on extrinsic parameters. However, prior works did not take the camera pose into account, which indicates the 3D convolution is performed based on an agnostic camera pose, entailing PA of the convolution scope.

Besides, the perspective transformation between the target 3D space and the 2D camera plane introduces Compu-

tation Imbalance (CI) on the 2D feature map, which is also demonstrated in Fig. 2 (b) and (c). Specifically, the convolved positions in the target 3D space, when projected on the 2D pixels, distributes quite sparse on the close scenes while dense on the far scenes. Such sparse computation allocation can hardly capture a comprehensive structural representation, from the rich details of structure or texture which usually exists in the 2D pixels projected from the close scenes, such as the red shelf in Fig. 2 (b) and (c).

Based on the three ambiguities the computation imbalance noticed, we devise a novel framework named NDC-Scene. Concretely, to alleviate FSA and FDA, the 3D feature maps are directly recovered in the NDC space, which strictly aligns with the image in hight and width, and extened in the depth-wise dimension. This methodology enables the implicit learning of precise occupancy and semantics among voxels, circumventing any erroneous inferences that may arise from 2D projection semantics. Furthermore, to address issues pertaining to PA and CI, we shift the majority computation units from the target 3D space to the NDC space. Extensive experiments on large-scale outdoor and indoor datasets demonstrate the superiority of our method to the existing state-of-the-art methods. The contributions can be summarized as follows:

- According to the critical problems we noticed in the existing methods, we propose a novel method based on Normalized Device Coordinates (NDC) space, which is proved to be a better space to put the majority 3D computation units than the target 3D space.

- In conjunction with the aforementioned camera space prediction, a pioneering depth-adaptive dual decoder is introduced to jointly upsample both 3D and 2D features and integrate them, thereby attaining more resilient representations.

- Experimental results demonstrate that the proposed method outperforms state-of-the-art monocular semantic scene completion methods significantly on both outdoor and indoor datasets.

2. Related Works

Single-View 3D Reconstruction infers the object-level or scene-level 3D geometry from a single RGB image. Most existing works focus on the reconstruction of a single object, which exploits encoder-decoder structures to learn explicit [44, 15, 11, 14, 42, 1, 25, 45, 47, 3, 32] or implicit [37, 28, 30, 31, 40, 41] representations of 3D objects and reconstruct the object's volumetric or surface geometry. A series of works [19, 16, 17, 49] extend this single-object 3D reconstruction to multi-object scenarios by reconstructing the instances detected in the image separately in a two-stage manner. For scene-level reconstruc-
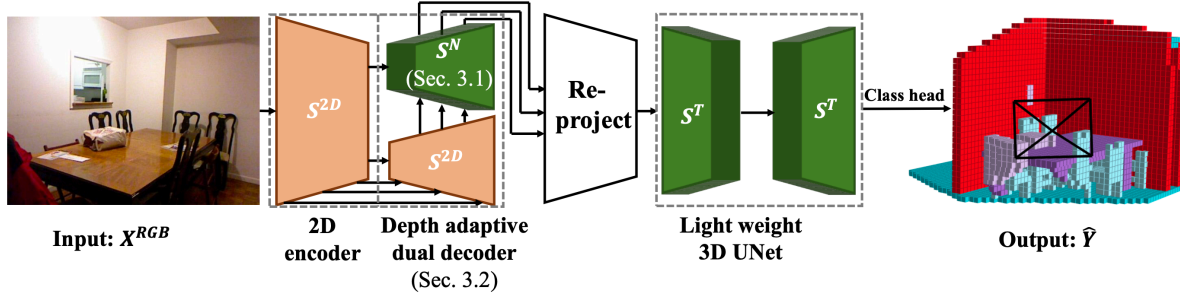
Figure 3: **NDC-Scene framework.** We first exploit an 2D image encoder to produce multi-scale 2D feature maps, followed by our Depth-Adaptive Dual Decoder (DADD Sec. 3.2) to restore the 3D feature map in S^N (Sec. 3.1), which is further re-projected to the target space S^T to predict the SSC result via a light-weight 3D UNet and a class head.

tion, [21, 18, 39, 29] combine the estimations of overall layout and objects to obtain a sparse holistic 3D reconstruction of the scene. [13] lifts the features of 2D panoptic segmentation to 3D to obtain the dense estimation of indoor scenes. However, most existing methods still cannot perform dense reconstruction robustly in various type of scenes. Although recent work [7] achieves dense semantic reconstruction of both indoor and outdoor scenes, it suffers from problems of PA and CI that limits its performance and robustness. In contrast, our method transfer most 3D computation units to a proposed normalized device coordinates space to avoid PA and CI, thus achieves better dense reconstruction in both indoor and outdoor scenes.

3D Semantic Scene Completion first defined in SSC-Net [38], aims to jointly infer scene geometry and semantics given incomplete visual observations. Previous works [27, 48, 23, 50, 22, 8, 6] have extensively studied SSC for indoor small-scale scenes and achieved satisfactory results. With the emergence of large-scale outdoor scene datasets and demands in autonomous driving, a series of works [34, 46, 33, 9] focus on the semantic completion of outdoor scenes, but such methods do not perform well in indoor scenes. At the same time, most existing works require RGB images along with additional geometric inputs, such as depth images [22], LiDAR point clouds [34], and Truncated Signed Distance Function (TSDF) [8, 6]. Vox-Former [24] proposes to leverage a pretrained depth estimator as 3D prior. But the requirement of geometric data limit the application of these methods. A notable exception is MonoScene [7], which first investigates monocular SSC that rely only on a single-view RGB image as input for scene completion. MonoScene proposes the Features Line of Sight Projection (FLoSP) to bridge 2D and 3D features, achieving competitive performance with models with additional geometric inputs or 3D inputs and generalizes well to different types of scenes. Nevertheless, the shared 2D features lifted to 3D rays via FLoSP results in both FSA and FDA, which limits its capacity in discerning the effective

patterns of depth and density. To improve the efficiency and performance, our method use a depth-adaptive dual decoder to restore the voxel feature on different depths in a more robust way, empowering it a strong representation of the occupancy and semantics among all depths.

3. Methodology

We consider a monocular 3D Semantic Scene Completion (SSC) task, which targets at predicting voxel-wise occupancy and semantics. Specifically, this task takes a single RGB image X^{RGB} as input and predicts volumetric labels $\hat{Y}$ in a target 3D space S^T with a shape (H^T, W^T, D^T). The labels $\hat{Y} \in C^{H^T \times W^T \times D^T}$ are divided in $M+1$ categories $C = \{c_0, c_1, ..., c_M\}$, with c_0 denoting the free voxel and $\{c_1, ..., c_M\}$ being the semantic categories.

Overview As mentioned above, current works in Monocular [7] SSC exploit Features Line of Sight Projection(FLoSP) to project 2D features to the target 3D space S^T, thus introduces the ambiguity of both the size and the depth in the projected 3D feature. Further, the positions of the 3D convolution operation in S^T, when projected on the 2D feature map, suffers from the problems of pose ambiguity and imbalanced computation allocation.

To tackle these problems, the proposed method extends the 2D feature map X^{2D} directly to the Normalized Device Coordinates space (NDC) S^N via reconstructing the depth dimensional with deconvoluion operations (Sec. 3.1). Such progressive reconstructed 3D features, compared with the shared 2D features projected along the camera ray in the FLoSP way, has the capacity of implicitly learning the density and depth of objects [2], and thus relieves both FSA anf FDA. Additionally, in S^N, the 2D projections of the 3D convolutions are evenly allocated. Such allocation has a stronger ability in capturing the structural representation of the rich details in the close scenes than the allocation of S^T. Finally, as S^N is invariant to the camera pose, the 3D convolution kernel in S^N has fixed offset, i.e., it has a con-
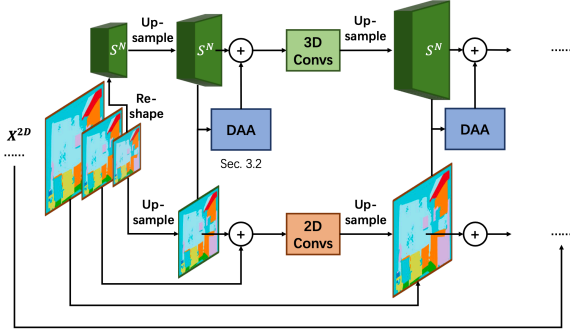
Figure 4: **Depth-adaptive dual decoder**. We infer the initial 3D feature map via a simple reshaping operation, and take the final feature map from the 2D image encoder as the the initial 2D feature map. In each decoder layer, the 2D features and 3D features are first upsampled by a scale factor 2 to $\boldsymbol{X}^{2D}_{s/2}$ and $\boldsymbol{X}^{N}_{s/2}$. Then $\boldsymbol{X}^{2D}_{s/2}$ is fused into $\boldsymbol{X}^{N}_{s/2}$ via DAA (Sec. 3.2), and a residual 2D feature map is combined to $\boldsymbol{X}^{2D}_{s/2}$. Finally, $\boldsymbol{X}^{2D}_{s/2}$ and $\boldsymbol{X}^{N}_{s/2}$ are processed by 3D and 2D convolution units respectively.

sistent semantic representation among all camera poses and the pose ambiguity is avoided in S^N.

Besides, we design a depth adaptive dual decoder (Sec. 3.2) to simultaneously upsample 3D feature maps in S^N and 2D feature maps derived from a 2D image encoder, respectively in two branches, as well as fuse them in each decoder layer to achieve more robust 3D semantic representations.

The pipeline of the proposed method is plotted in Fig. 3. At first, the input RGB image is encoded by a pre-trained 2D image encoder to generate multi-scale 2D feature maps. Afterwards, the proposed dual decoder is responsible reconstruct the 3D feature map in S^N, which is further reprojected into the target space S^T. Finally, a 3D UNet in S^T processes the projected 3D features and predicts the completion result. In the proposed method, the 3D UNet in S^T is quite light-weight, consisting of only downsample and upsample modules. While most 3D computation units is transferred to the 3D branch in the proposed dual decoder, which restore the 3D feature maps in S^N. We prove in Sec. 4 that moving the majority of 3D computation cost from the target 3D space S^T to the proposed S^N brings obvious performance gains.

3.1. Normalized Device Coordinates Space

Feature Ambiguity   Since the monocular SSC task [7] assumes a RGB image from only a single view of point as input, it is impossible to back-project 2D features to their exact 3D correspondences due to lack of the guidance of depth. Current works [7] exploits FLoSP to project 2D features to all possible locations in the target 3D space S^T along their lines of sight. This practice, although shown to be effective, results in the ambiguity in the projected 3D features. This ambiguity can be categorized as Feature-Size

Ambiguity (FSA) and Feature-Depth Ambiguity (FDA). As for FSA, the 2D image features are spread into larger space as the depth gets larger, as a result, such distribution with agnostic density makes it hard for the convolution kernel to determine the effective pattern. In a similar way, for depth, the 3D networks responsible for the SSC in S^T can hardly distinguish the shared 2D features among all possible depths to discern the reasonable positions. An intuitive demonstration of feature ambiguity is shown in Fig. 1 (b).

Also, we find two more imperceptible drawbacks existing in the mainstream monocular SSC works [7]. To begin with, we first define the 2D space of the pixels in X^{RGB} as S^{2D}, and the coordinates in S^{2D} as $\boldsymbol{p}^{2D}$, formally:

$$S^{2D} = \left[0, W^{2D}\right] \times \left[0, H^{2D}\right], \qquad (1)$$
$$\boldsymbol{p}^{2D}_{i,j} = \left(x^{2D}_{i,j}, y^{2D}_{i,j}\right) \in S^{2D}. \qquad (2)$$

In common practice, a 2D feature map in the 2D space S^{2D} is generated from a UNet structure, and then projected to the target 3D space, with the coordinates $\boldsymbol{p}^T$ and space S^T represented as:

$$S^T = \left[0, W^T\right] \times \left[0, H^T\right] \times \left[0, D^T\right], \qquad (3)$$
$$\boldsymbol{p}^T_{i,j,k} = \left(x^T_{i,j,k}, y^T_{i,j,k}, z^T_{i,j,k}\right) \in S^T. \qquad (4)$$

The coordinates in S^T and S^{2D} are related via an affine transformation decided by the camera pose, formally, the extrinsic parameters $\boldsymbol{a}^R, \boldsymbol{b}^R$, followed by a perspective transformation, with parameters $\boldsymbol{f}, \boldsymbol{c}$, formally:

$$\left(x^R_{i,j,k}, y^R_{i,j,k}, d^R_{i,j,k}\right) = \boldsymbol{a}^R \boldsymbol{p}^T_{i,j,k} + \boldsymbol{b}^R, \qquad (5)$$
$$\boldsymbol{p}^{2D}_{i',j'} = \boldsymbol{f} \cdot \left(x^R_{i,j,k}, y^R_{i,j,k}\right) / d^R_{i,j,k} + \boldsymbol{c}. \quad (6)$$

The 3D point $\boldsymbol{p}^T_{i,j,k}$ is first projected to the camera coordinate system via the affine transformation, with the projected coordinate $\left(x^R_{i,j,k}, y^R_{i,j,k}, d^R_{i,j,k}\right)$. This coordinate is further projected to $\boldsymbol{p}^{2D}_{i',j'}$ in the 2D image space via the perspective transformation, with focal length f and image center c. We find that both the two transformations introduce discrepancy between the convolution in the 3D space S^T and the original pixel arrangement in S^{2D}.

Pose Ambiguity   First, as a scene can be projected to multiple possible target 3D spaces for the SSC task, the affine transformation $\boldsymbol{a}^R, \boldsymbol{b}^R$ also has myriad possibilities. Currently, dataset providers usually selects $\boldsymbol{a}^R, \boldsymbol{b}^R$ following hand-crafted heuristic strategy, e.g., making a border of S^T parallel to a border of the room, such as a wall. According to Eq. 5, the affine transformation decides the correspondence between the 2D coordinate $\boldsymbol{p}^{2D}_{i,j}$ and the 3D coordinate $\boldsymbol{p}^T_{i,j,k}$ in S^T, and thus the relative positions between the 2D projections of two 3D coordinates. However, the
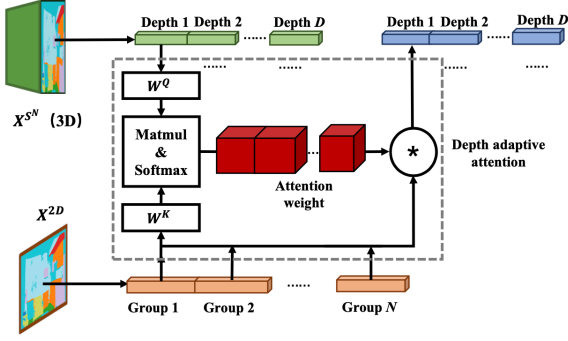
Figure 5: **Depth-adaptive attention**. We infer the attention matrix $\boldsymbol{a}$ via the inner-production between the 3D query feature and the 2D key feature on each group. We omit the value projection for computation reduction.

selection strategy of $\boldsymbol{a}^R, \boldsymbol{b}^R$ is agnostic to the SSC model, which means the convolution operation in S^T is unaware of which positions in the 2D space it is really convolving on. To be specific, we consider a 3D convolution kernel $\boldsymbol{W} \in R^{(2K+1)^3 C_{in} C_{out}}$ which performs 3D convolution on a 3D position $\boldsymbol{p}_{i,j,k}^T$ in S^T, with the kernel size $2K+1$. Its output on position $\boldsymbol{p}_{i,j,k}^T$ can be represented by:

$$\boldsymbol{O}_{i,j,k} = \sum_{(i',j',k') \in I_r} \boldsymbol{W}_{i'-i,j'-j,k'-k} \boldsymbol{X}_{i',j',k'}^T, \quad (7)$$

$$\begin{aligned} I_{i,j,k}^K = \{i-K, ..., i, ..., i+K\} \\ \times \{j-K, ..., j, ..., j+K\} \\ \times \{k-K, ..., k, ..., k+K\}. \end{aligned} \quad (8)$$

Where $I_{i,j,k}^K$ is the convolution scope in S^T when the kernel convolves on $\boldsymbol{p}_{i,j,k}^T$ and $X^T \in R^{C_{in} \times H^T \times W^T \times D^T}$ is the 3D feature map in S^T. Since the convolved 3D feature $\boldsymbol{X}_{i',j',k'}^T$ is projected from the 2D feature map in S^{2D}, we further consider the back-projected 2D positions $\boldsymbol{p}_{I_{i,j,k}^K}^{2D}$ of the 3D positions $\boldsymbol{p}_{I_{i,j,k}^K}^T$ in this convolution scope, as well as the back-projected 2D position $\boldsymbol{p}_{i',j'}^{2D}$ of the 3D position $\boldsymbol{p}_{i,j,k}^T$. As shown in Fig. 2 (b) and (c), the relative position between $\boldsymbol{p}_{I_{i,j,k}^K}^{2D}$ and $\boldsymbol{p}_{i,j,k}^T$ is inconsistent between different selections of $\boldsymbol{a}^R, \boldsymbol{b}^R$, which implies that, given the same 2D feature map, 3D convolution in S^T can hardly learns a robust semantic representation among different choice of $\boldsymbol{a}^R, \boldsymbol{b}^R$, named as Pose Ambiguity (PA).

Imbalanced Computation Allocation Next, the perspective transformation $\boldsymbol{f}, \boldsymbol{c}$ introduces another problem between the 2D and 3D space, namely, the Imbalanced Computation Allocation. According to Eq. 5, $\boldsymbol{p}_{I_{i,j,k}^K}^{2D}$ distribute quite sparse on locations in close scenes due to the shallow depths, while very dense in the distant ones. Such imbal-

anced allocation is also shown in Fig. 2 (b) and (c). However, the raw vision information are uniformly distributing among the 2D pixels, hence we argue that a uniform allocation of computation over the 2D space is more robust than the shown one. More intuitively, we notice that the close objects, when projected on the 2D image, contains more detailed structure or texture than the far ones. It means a sparse computation allocation can hard capture a strong and comprehensive structural representation of the close objects. For example, in Fig. 2 (b) and (c), the close scene contains rich detail of a red shelf, while the far scene are mostly composed of a simple wall.

Normalized Device Coordinates Space To solve the three problems mentioned above, we propose a normalized device coordinates space S^N, with the coordinates $\boldsymbol{p}_{i,j,k}^C = (x_{i,j,k}, y_{i,j,k}, d_{i,j,k})$. S^N is derived via directly extending the location $\boldsymbol{p}_{i,j}^{2D} = (x_{i,j}, y_{i,j})$ in the 2D pixels space with an additional dimension $d_{i,j,k}$, which is the depth to the camera plane. In this way, the 3D convolution operation has a consistent scope among different choices of affine transformations, as well as evenly distributed computation allocation among the 2D space S^{2D}, as shown in Fig. 2 (a). It means that S^N avoids the pose ambiguity and imbalanced computation allocation. Further, as described in Sec. 3.2, the progressively restored information in the dimension of depth endows the 3D feature map in S^N a strong representation of the occupancy and semantics in different depths, especially compared to the shared feature among the sight of line in FLoSP. An intuitive comparison is provided in Fig. 1 (a) and (b).

3.2. Depth Adaptive Dual Decoder

We find that transferring the majority amount of 3D process from S^T to the proposed normalized device coordinates space S^N brings obvious performance gain, as illustrated in Sec. 4. To achieve a robust semantic representation in S^N, we propose a Depth Adaptive Dual Decoder (DADD). DADD simultaneously performs upsample on the 2D and 3D feature map, respectively in two branches of decoder layers, as well as fuse the 2D feature to 3D with a novel Depth Adaptive Attention (DAA) module.

Dual Decoder The input of DADD is a 2D feature map $\boldsymbol{X}_S^{2D} \in R^{C_{in} \times W^{2D} \times H^{2D}}$ generated from a 2D encoder, which has a downscale stride s. Then, to achieve the initial 3D feature map, we exploit a reshaping operation to divide the dimension of channel into D groups, each denoting an

Method	Input	SC IoU	ceiling (1.37%)	floor (17.58%)	wall (15.26%)	window (1.99%)	chair (3.01%)	bed (7.08%)	sofa (4.70%)	table (4.31%)	tvs (0.47%)	furniture (30.04%)	objects (14.19%)	mIoU
LMSCNet[rgb] [34]	Occ	33.93	4.49	88.41	4.63	0.25	3.94	32.03	15.44	6.57	0.02	14.51	4.39	15.88
AICNet[rgb] [22]	RGB & Depth	30.03	7.58	82.97	9.15	0.05	6.93	35.87	22.92	11.11	0.71	15.90	6.45	18.15
3DSketch[rgb] [8]	RGB & TSDF	38.64	8.53	90.45	9.94	5.67	10.64	42.29	29.21	13.88	9.38	23.83	8.19	22.91
MonoScene [7]	RGB	42.51	8.89	93.50	12.06	12.57	13.72	48.19	36.11	15.13	15.22	27.96	12.94	26.94
NDC-Scene(ours)	RGB	44.17	12.02	93.51	13.11	13.77	15.83	49.57	39.87	17.17	24.57	31.00	14.96	29.03

Table 1: **Quantitative comparsion** against RGB-inferred baselines and the state-of-the-art monocular SSC method on NYUv2 [37]. The notations Occ, Depth and TSDF denote the occupancy grid(3D), depth map(2D) and TSDF array(3D), which are the 3D input required by the SSC baselines. For a fair comparison, all the three input are converted from the depth map predicted by a pretrained depth predictor [5]
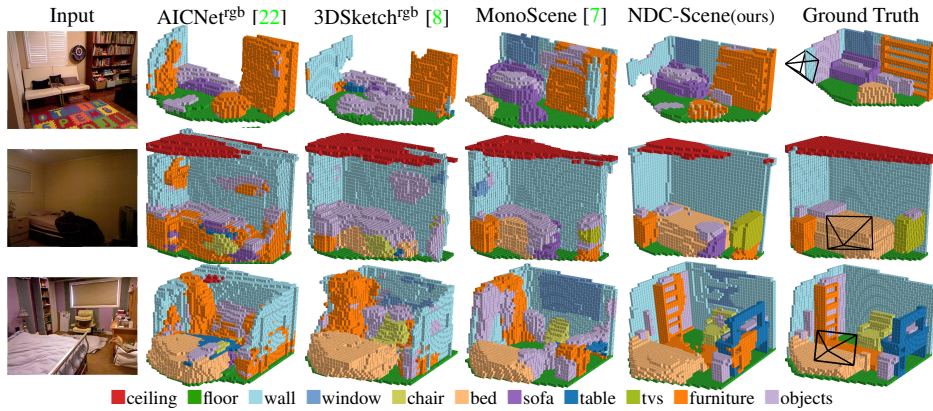

ceiling floor wall window chair bed sofa table tvs furniture objects

Figure 6: **Qualitative results on NYUv2 [37] (test set).** From left to right: (a) RGB input, (b) results of AICNet[rgb] [22], (c) results of 3DSketch[rgb] [8], (d) results of MonoScene [7], (e) ours results. NDC-Scene achieve higher voxel-level accuracy and better semantic predictions on NYUv2 (test set) compared with existing SSC baselines.

individual depth, formally:

$$X_S^{2D} \in R^{C_{in} \times W^{2D} \times H^{2D}}$$
$$\xrightarrow{reshape}$$
$$X_s^{3D} \in R^{C_{in}/D \times W^{2D} \times H^{2D} \times D}. \quad (9)$$

Afterwards, in each decoder layer, we first transform the 2D feature map X_S^{2D} to a larger resolution $X_{s/2}^{2D}$ with scale factor 2, following the common practice [36, 7] in the decoder of the 2D UNet, which upsamples X_S^{2D} and add it to the residual feature map with the same resolution generated in the corresponding 2D encoder layer, followed by several 2D convolution units. Then, the 3D feature map X_s^{3D} is also upsampled to $X_{s/2}^{3D}$. Afterwards, the upscaled X_S^{2D} is fused into $X_{s/2}^{3D}$ via the proposed DAA module, followed by several 3D convolution operations. We demonstrate more intuitive details in Fig. 4.

Depth Adaptive Attention We assume that, with sufficiently large receptive field, a 2D feature at a 2D position $p_{i,j}^{2D}$ in X_S^{2D} can aggregate the context information to im-

plicitly infers both the surface and behind scenes at $p_{i,j}^{2D}$. As fully verified in objective detection [12, 26], objects in different depth projected in the same 2D location, can be encoded in different channels of the detection head. In this paper, we also assume that the 3D semantic scene in different depth projected at $p_{i,j}^{2D}$ can be encoded in different channel groups of the 2D feature $X_{i,j}^{2D}$. From this view of point, we propose depth adaptive attention to facilitate 3D features $X_{i,j,k}^T$ in each depth-of-field to flexibly decide, which channel group of the 2D feature $X_{i,j}^{2D}$ in the position $p_{i,j}^{2D}$ it is projected on, is most helpful to restore a robust representation of its depth-of-field. Formally, we divide the 2D feature $X_{i,j}^{2D}$ with C_{in} channels to G groups, each with C_{in}/G channels. The depth adaptive attention of $X_{i,j,k}^T$ on $X_{i,j}^{2D}$ is represented as follows:

$$A_{i,j,k} = Softmax\left(a_{i,j,k}\right), \quad (10)$$
$$a_{i,j,k}^g = \left(W^Q X_{i,j,k}^T\right)' \cdot W^K X_{i,j,g}^{2D}, \quad (11)$$
$$X_{i,j,k}^T = A_{i,j,k} X_{i,j}^{2D}. \quad (12)$$

Method	SSC Input	SC IoU	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-grnd (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle (0.03%)	other-veh. (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (%)	person (0.07%)	bicyclist (0.07%)	motorcyclist. (0.05%)	fence (3.90%)	pole (0.29%)	traf.-sign (0.08%)	mIoU
LMSCNet[rgb] [34]	Occ	28.61	40.68	18.22	4.38	0.00	10.31	18.33	0.00	0.00	0.00	0.00	13.66	0.02	20.54	0.00	0.00	0.00	1.21	0.00	0.00	6.70
3DSketch[rgb] [8]	RGB & TSDF	33.30	41.32	21.63	0.00	0.00	14.81	18.59	0.00	0.00	0.00	0.00	19.09	0.00	26.40	0.00	0.00	0.00	0.73	0.00	0.00	7.50
AICNet[rgb] [22]	RGB & Depth	29.59	43.55	20.55	11.97	0.07	12.94	14.71	4.53	0.00	0.00	0.00	15.37	2.90	28.71	0.00	0.00	0.00	2.52	0.06	0.00	8.31
MonoScene [7]	RGB	37.12	57.47	27.05	15.72	0.87	14.24	23.55	7.83	0.20	0.77	3.59	18.12	2.57	30.76	1.79	1.03	0.00	6.39	4.11	2.48	11.50
NDC-Scene(ours)	RGB	**37.24**	**59.20**	**28.24**	**21.42**	**1.67**	**14.94**	**26.26**	**14.75**	**1.67**	**2.37**	**7.73**	19.09	**3.51**	**31.04**	**3.60**	**2.74**	0.00	**6.65**	**4.53**	**2.73**	**12.70**

Table 2: **Quantitative comparsion** against RGB-inferred baselines and the state-of-the-art monocular SSC method on SemanticKITTI [4]. The notations Occ, Depth and TSDF denote the occupancy grid(3D), depth map(2D) and TSDF array(3D), which are the 3D input required by the SSC baselines. For a fair comparison, all the three input are converted from the depth map predicted by a pretrained depth predictor [5]
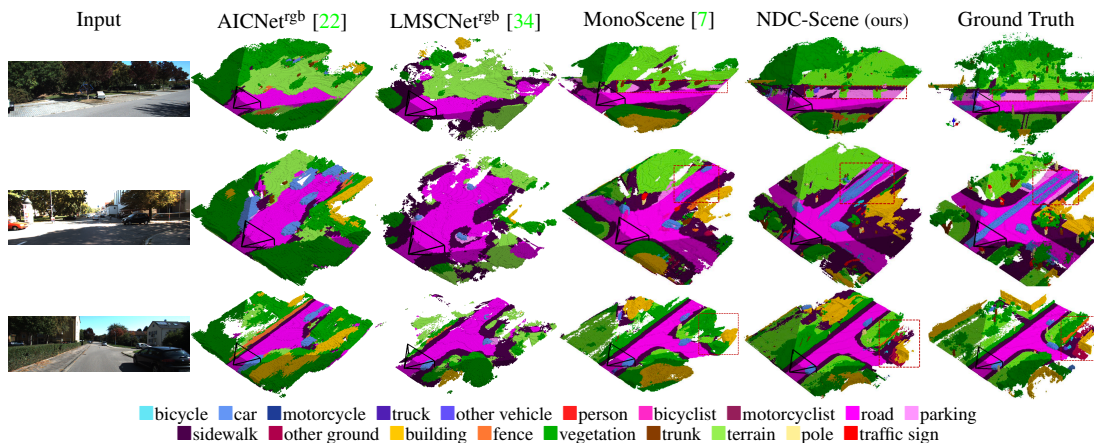

Figure 7: **Qualitative results on SemanticKITTI [4] (val set).** From left to right: (a) RGB input, (b) results of AICNet[rgb] [22], (c) results of LMSCNet[rgb] [34], (d) results of MonoScene [7], (e) ours results. NDC-Scene achieve higher voxel-level accuracy and better semantic predictions on SemanticKITTI (val set) compared with existing SSC baselines.

W^Q and W^K are two projection matrices to calculate the similarity vector a. This design is mostly inspired by the attention module in [10], which omits the value projection W^V to save the computational cost. The detail structure is shown in Fig. 5.

4. Experiment

We evaluate NDC-Scene on the didactic real-world indoor dataset NYUv2 [37] and outdoor SemanticKITTI [4]. We compare NDC-Scene with both state-of-the-art monocular SSC baselines and several adapted SSC baselines which requires additional depth information. Both datasets contains labeled ground truth in the target space and camera extrinsics.

NYUv2 Dataset [37] consists of 1449 scenes captured using the Kinect camera, which are represented as $240 \times 144 \times 240$ voxel grids annotated with 13 classes (11 semantic, one free, and one unknown). The shape of the RGB images is 640×480. Following previous works [7], we split

the dataset with 795 training samples and 654 test samples.

SemanticKITTI Dataset [4] is a large-scale outdoor 3D scene semantic complement dataset, which contains Lidar scans represented as $256 \times 256 \times 32$ grids of 0.2m voxels. The dataset includes 21 classes, including 19 semantic labels, one free label, and one unknown label. We utilize the RGB images of cam-2 with a resolution of 1226×370 and left cropped to 1220×370. And follow the official train / val splits consisting 3834 and 815 samples, respectively. Following [7], we evaluate our models at full scale (*i.e.* 1:1). Our main results and ablations are performed using the offline validation set for convenience, and the results on the hidden test set on the online server are also provided in the supplementary materials.

Metrics For the scene completion (SC) metric, we present the intersection over union (IoU) of occupied voxels regardless of their semantic classification following common practice. For the semantic scene completion (SSC) metric, we report the mean IoU (mIoU) across all semantic

Methods	w/o FA	w/o CI	w/ DA	NYUv2		SemanticKITTI	
				IoU ↑	mIoU ↑	IoU ↑	mIoU ↑
Ours	✓	✓	✓	**44.17**	**29.03**	**37.24**	**12.70**
NDC-FA	×	✓	✓	42.96 (-1.21)	27.69 (-1.34)	37.15 (-0.09)	12.03 (-0.67)
NDC-CI	✓	×	✓	43.72 (-0.45)	28.10 (-0.93)	37.20 (-0.04)	12.26 (-0.44)
NDC-NF	✓	✓	×	43.52 (-0.65)	28.22 (-0.81)	37.18 (-0.06)	11.22 (-0.48)
MonoScene	×	×	×	42.51 (-1.66)	26.94 (-2.09)	37.12 (-0.12)	11.50 (-1.20)

Table 3: **Components ablation.** All of our components boost performance consistently on NYUv2 [37] and SemanticKITTI [4].

categories. Note that the training and evaluation processes differ for indoor and outdoor settings because of the different depth and sparsity of the LiDAR data. To account for both scenarios, we use the more challenging evaluation metrics for all voxels, and we follow [7] to report the baseline results under the consistent metric.

Implementation Details   Our experiments are conducted on 2 NVIDIA Tesla V100 SXM2 GPUs. Specifically, we exploit DDR module [22] and 3D deconvolution layer as the 3D computation unit and the upsample operation in DDAD. As for the 3D UNet, the downsample and upsample operation are instantiated as 3D convolution and deconvolution layer, both with stride 2. The initial 2D/3D feature maps in DDAD are 15x20/15x20x16 and 39x12/39x12x32, respectively for NYUv2 and SemanticKITTI, while the output 3D feature maps of DADD are 60x80x64 and 156x48x128. We follow the loss functions in [7]. The group number g of DAA is 8.

4.1. Performance

We compare our proposed NDC-Scene with existing strong SSC baselines designed for indoor (3DSketch [8], and AICNet [22]) or outdoor (LMSCNet [34]) scenarios. We also compare NDC-Scene with MonoScene [7], which is the best RGB-only SSC method. Note that for the methods with more than RGB inputs, we follow [7] to adapt their results to RGB-only inputs.

NYUv2   Tab. 1 presents the performance of NDC-Scene on NYUv2 [37] compared with other SSC methods, which outperforms all other methods by a considerable margin. Compared with previous state-of-the-art MonoScene [7], our method obtains better results not only on both IoU and mIoU but also on all categories, demonstrating the effectiveness and robustness of our proposed architecture.

SemanticKITTI   We compare the results under the outdoor scenarios on SemanticKITTI [4] in Tab. 2. For all categories including small (*e.g.* car, bicycle, and person) and large(*e.g.* building, truck, and road) semantics, our method beat the baselines significantly, showing that our method can adapt to different scenarios.
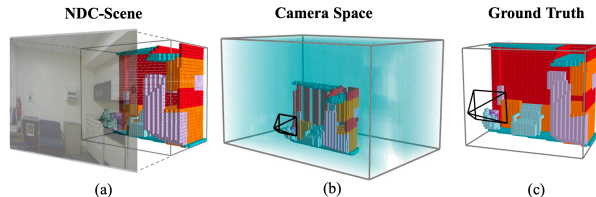

Figure 8: **Qualitative comparison** of the normalized device coordinates space and the camera space. In (b), the voxels of S^R that are close to the camera are mostly out of the image when projected on the camera plane. Thus the in-image voxels are quite sparse. In contrast, in (a) the voxels with different depths are all uniformly projected on the camera plane.

Qualitative performance   In Fig. 6, we present visualization results to qualitatively evaluate the effectiveness of our NDC-Scene on NYUv2 [37] dataset. We can see that the proposed NDC-Scene has the ability to handle a wide range of objects with diverse shapes, resulting in more precise scene layout and instance-level information than other SSC baselines. Fig. 7 illustrates the qualitative results on SemanticKITTI [4]. Our NDC-Scene demonstrates satisfactory performance in discerning the accurate depth range of objects like terrains, trunks and cars, as highlighted in the red boxes. This improvement holds great significance in large-scale outdoor scenarios and serves as compelling evidence that we have greatly relieved the issue of Feature Ambiguity exposed by MonoScene [7].

4.2. Ablation Study

We design comprehensive experiments to verify the capacity of the proposed method in solving the identified problems of feature ambiguity, pose ambiguity and computation imbalance.

Feature Ambiguity   To evaluate NDC-Scene's ability in relieving the feature ambiguity problem, we replace the proposed dual decoder with a heavy 3D UNet with the same amount of FLoSP, the input features of which are lifted via FLoSP. This variant loses the ability of implicitly restoring the semantics among different depths, named as NDC-FA. For the detailed architecture of NDC-FA, please refer to the supplementary. Tab. 3 shows that, with the introduced feature ambiguity, the performance in both geometry ([-1.21,

	Ours(NYUv2)		MonoScene(NYUv2)	
	IoU ↑	mIoU ↑	IoU ↑	mIoU ↑
$\theta = 0°$	**44.17**	**29.03**	**42.51**	**26.94**
$\theta = 5°$	42.88 (-1.29)	28.28 (-0.75)	38.99 (-3.52)	23.20 (-3.74)
$\theta = 10°$	39.07 (-5.10)	24.64 (-4.39)	33.52 (-8.99)	20.15 (-6.79)
$\theta = 15°$	36.74 (-7.43)	22.39 (-6.64)	30.05(-12.46)	16.71 (-10.23)

Table 4: **Ablation study** for the robustness to pose ambiguity on NYUv2 [37].

-0.09] IoU) and semantics ([-1.34, -0.67] mIoU) degrades a lot.

Computation Imbalance Also, NDC-CI is designed to verify the capacity of NDC-Scene in solving the computation imbalance problem, Where the 3D feature map in DADD is replaced with a camera space S^R, with coordinates $\boldsymbol{p}_{i,j,k}^R = \left(x_{i,j,k}^R, y_{i,j,k}^R, d_{i,j,k}^R \right)$ uniformly distributing in S^R. As the 2D pixels $\boldsymbol{p}_{i,j}^{2D}$ and $\boldsymbol{p}_{i,j,k}^R$ are also related by Eq. 6 in the same way of $\boldsymbol{p}_{i,j}^{2D}$ and $\boldsymbol{p}_{i,j,k}^T$, 3D convolution in S^R also suffers from the CI problem. A intuitive comparison between S^R and the proposed S^N is illustrated in Fig. 8. We notice that NDC-CI degrades much more in semantics ([-0.93, -0.44] mIoU) than in geometry ([-0.45, -0.04] IoU), as shown in Tab. 3.

Depth Adaptive Attention We compare DAA with a naive fusion approach which direct adds the 2D feature on $\boldsymbol{p}_{i,j}^{2D}$ to all 3D features that projects at $\boldsymbol{p}_{i,j}^{2D}$, named as NDC-NF. Accoring to Tab. 3, DAA compensates for 30% of the performance gain compared to MonoScene.

Pose Ambiguity For PA, we randomly rotate target space as well as the SSC label with an angle uniformly sampled in $[0, \theta]$, with θ in $[5°, 10°, 15°]$, and compare the performance degradation of NDC-Scene and MonoScene to validate the robustness of NDC-Scene in the choice of extrinsic camera parameters. As revealed in Tab. 4, the performance of NDC-Scene degrades much slower than MonoScene as the increase of θ, which verifies the strong ability of NDC-Scene in handling pose ambiguity.

5. Conclusion

To conclude, our study comprehensively explores the critical challenges encountered by the present state-of-the-art techniques in monocular 3D semantic scene completion. To overcome these challenges, the proposed method introduces a novel Normalized Device Coordinates (NDC) space predictor technique, which effectively extends the 2D feature map to a 3D space by progressively restoring the dimension of depth with deconvolution operations. By transferring the majority of computation from the target 3D space to the proposed normalized device coordinates space, the proposed approach leads to enhanced performance in monocular SSC tasks. Furthermore, the study proposes the use of a Depth-Adaptive Dual Decoder, which facilitates the simultaneous upsampling and fusion of the 2D and 3D feature maps, thereby improving overall performance.

6. Acknowledgements

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *ICML*, 2018.

[2] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018.

[3] Zhiqin Chen andrea Tagliasacchi and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *CVPR*, 2020.

[4] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019.

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[6] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2021.

[7] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022.

[8] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4193–4202, 2020.

[9] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *CoRL*, 2021.

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul

Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[11] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016.

[12] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12214–12223, 2020.

[13] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. In *NeurIPS*, 2021.

[14] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017.

[15] Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016.

[16] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *ICCV*, 2019.

[17] Alexander Grabner, Peter M Roth, and Vincent Lepetit. Location field descriptors: Single image 3d model retrieval in the wild. In *3DV*, 2019.

[18] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *ECCV*, 2018.

[19] Hamid Izadinia, Qi Shan, and Steven M. Seitz. Im2cad. In *CVPR*, 2017.

[20] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view pointnet for 3d scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[21] Chen-Yu Lee, Vijay Badrinarayanan, and Tomasz Malisiewicz andrew Rabinovich. Roomnet: End-to-end room layout estimation. In *ICCV*, 2017.

[22] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3351–3359, 2020.

[23] 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*, 2019.

[24] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. *arXiv preprint arXiv:2302.12251*, 2023.

[25] Yiyi Liao, Simon Donné, and Andreas Geiger. Deep marching cubes: Learning explicit surface representations. In *CVPR*, 2018.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[27] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. *Advances in Neural Information Processing Systems*, 31, 2018.

[28] Lars Mescheder, Michael Oechsle, Michael Niemeyer, and Sebastian Nowozin andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019.

[29] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *CVPR*, 2020.

[30] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.

[31] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, and Marc Pollefeys andreas Geiger. Convolutional occupancy networks. In *ECCV*, 2020.

[32] Jingtan Piao, Keqiang Sun, Quan Wang, Kwan-Yee Lin, and Hongsheng Li. Inverting generative adversarial renderer for face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15628, 2021.

[33] Christoph B Rist, David Emmerichs, Markus Enzweiler, , and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE TPAMI*, 2021.

[34] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020.

[35] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3d semantic scene completion: A survey. *International Journal of Computer Vision*, 130(8):1978–2005, 2022.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[37] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012.

[38] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017.

[39] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *CVPR*, 2019.

[40] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and Hongsheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. In *Advances in Neural Information Processing Systems*, 2022.

[41] Keqiang Sun, Shangzhe Wu, Ning Zhang, Zhaoyang Huang, Quan Wang, and Hongsheng Li. Cgof++: Controllable 3d face synthesis with conditional generative occupancy fields. *arXiv preprint arXiv:2211.13251*, 2022.

[42] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018.

[43] Christian Wojek, Stefan Walk, Stefan Roth, and Bernt Schiele. Monocular 3d scene understanding with explicit occlusion reasoning. In *CVPR 2011*, pages 1993–2000. IEEE, 2011.

[44] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NeurIPS*, 2016.

[45] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *IJCV*, 2020.

[46] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3101–3109, 2021.

[47] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *CVPR*, 2020.

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

[49] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020.

[50] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *CVPR*, 2019.