IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 February 2024, accepted 24 February 2024, date of publication 26 February 2024, date of current version 4 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3370844

RESEARCH ARTICLE

Object-Aware Semantic Scene Completion Through Attention-Based Feature Fusion and Voxel-Points Representation

YUBIN MIAO, JUNKANG WAN, JUNJIE LUO, HANG WU, AND RUOCHONG FU
School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Yubin Miao (ybmiao@sjtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 51975361.

ABSTRACT Semantic scene completion is a computer vision technique that combines semantic segmentation and shape completion. Its purpose is to infer a complete 3D scene with semantic information from single-view RGB-D images. In recent years, some methods have adopted the voxel-points-based approach, converting voxelized scenes into point clouds to reduce the computational cost associated with 3D convolutions. However, majority of such methods do not fully consider the geometric details of the objects in the scene. In this paper, we propose ASPNet (Attention-based Semantic Point Completion Network), a two-branch semantic scene completion algorithm that combines scene-level completion and object refinement. In the scene level completion branch, we design the SPT (Semantic-based Point Transformer) module, which introduces semantic information into the traditional Point Transformer layer to realize the feature aggregation of neighboring keypoints of the same category. Using the object detection module and the object refinement module, ASPNet refines the rough semantic complementation results obtained from direct coding and decoding of RGB-D inputs. The quantitative results show that ASPNet has much less computational overhead than the 3D convolution-based semantic scene completion algorithm, while the reconstruction results have more geometric details.

INDEX TERMS Semantic scene completion, semantic segmentation, 3D scene reconstruction, deep learning, point transformer, point cloud.

I. INTRODUCTION

For intelligent devices, the ability to infer the complete scene from a single perspective RGB-D image is significant, which can be widely applied in autonomous driving, augmented reality, and robotics technology [1]. Therefore, computer vision researchers are committed to studying how to make machines possess such capabilities. To address this issue, Semantic Scene Completion (SSC) has been proposed, aiming to instruct machines on how to understand the 3D world from static depth and/or RGB images. This task has two coupled objectives: one is to complete the 3D scene, aiming to infer the volume occupancy of the scene; The other is to label 3D scene, which requires wise prediction of semantic label voxels. Due to the presence of occlusion in the real world, there are significant changes in the shape, layout, and visibility of objects. Therefore, the main challenge is how to model the 3D context to effectively learn each voxel [2].

At present, 3D convolutional neural networks (3D-CNNs) are a commonly used method in the task of SSC, which uses a classic ‘‘encoder-decoder’’ structure composed of a series of 3D convolutional layers to achieve functional mapping from feature volumes to semantic volumes. However, recent research results [3], [4] have shown that due to the sparsity of 3D data, more than 85% of voxels in front of visible surfaces can be directly labeled as empty (i.e., most voxels in 3D scenes are empty), leaving less than 15% of voxels that require predicting semantic labels [5]. Therefore, mainstream methods (based on convolution for SSC) will consume a large amount of computational resources on these empty

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

voxels. To tackle this problem, some methods [3], [5] convert effective voxel data into points and introduce them into SSC tasks. These algorithms are implemented using only a few 3D convolutional layers, greatly reducing the computational cost of pure 3D convolution. However, due to the insufficient consideration of the features of these ''voxel points'' and their differences from surface point clouds, there is still room for improvement in point-based SSC methods.

It is worth noting that 2D semantic segmentation (SS) can predict class labels from RGB images. Although the dimensions of 2D SS and 3D SSC are inconsistent, they share some important features that enable them to match each other [6]. For example, they generate the same element semantic labels. After 2D-3D projection, the two-dimensional semantic segmentation results are projected onto the three-dimensional feature volume. In addition, RGB images have rich texture features that can supplement the edge contours of various objects in 3D scenes. Intuitively speaking, the dense prediction of 2D SS can compensate for the sparsity of 3D SSC, and the complete shape/layout of 3D scenes can help distinguish indivisible 2D regions [6]. In the 3D scene completion method, an intuitive and direct concept, namely the ''detection and completion'' strategy, has been proven to be effective [7], [18]. Detecting objects in 3D space [29], [30] can provide prior information for SSC tasks. Through this method, not only can the boundaries between different objects be more clearly defined, but also the internal shape and structure of these objects can be more accurately predicted.

In this paper, we proposed a dual branch network that takes a pair of single view RGB-D images as input. One is the SSC branch based on voxel-points, and the other is the pre-trained 3D object detection branch. The two branches are fused by the geometric refinement module to generate the final refined semantic completion result. Specifically, the first branch consists of two sub modules: a pre-trained 2D semantic segmentation module and a point-based SSC module. Firstly, our 2D semantic map will be projected into the 3D semantic volume. Then, we extract effective voxels within two volumes and convert them into point clouds for subsequent SSC modules. The backbone of the SSC module uses PointNet++ [35] as the basic framework, and adds a Point Transformer layer for feature fusion after each Set Abstraction (SA) layer and Adaptive Feature Propagation (A-FP) layer. After completing the shallow surface feature extraction for the surface point cloud as well as for the effective point cloud, the Surface-Attention module is utilized to realize the supplementation of the features extracted from the effective point cloud for each layer. In addition, after completing the feature extraction, the Feature Propagation (FP) layer is replaced with an adaptive FP layer (A-FP) based on the attention mechanism during the up-sampling process. At the same time, we replace the Point Transformer module, which is located in the last layer, with a newly designed Semantic-based Point Transformer (SPT) layer to efficiently aggregate the category identities. The second branch locates each indoor object and determines the corresponding 3D object detection box for each object in the space. We crop and refine the voxels within the bounding box, and then use the geometric refinement module to refine and correct the voxels located in the target detection box, thereby achieving accurate semantic reconstruction.

Compared to existing methods, we design the Semantic-based Point Transformer (SPT) module introduces a semantic aggregation mechanism into the Point Transformer module, enabling the aggregation of features for adjacent points of the same category. The Surface-Attention module supplements internal voxel points with features from feature-rich surface voxel points in order to address the feature imbalance among voxel points.

In summary, contributions of our work are as follows:

1) We introduce a novel dual-branch structure (ASP-Net) for semantic scene completion, which combines point-based SSC and object detection to reconstruct accurate indoor scenes. The structure uses pre-trained bounding boxes as guidance for refinement modules to reconstruct more geometric details from rough completion results.
2) We designed a Semantic based Point Transformer (SPT) layer that integrates binary classification judgment. It improves the similarity of local features of similar voxel-points in high-dimensional space, thereby avoiding the occurrence of outliers and ''noise''.
3) Our method adopts a main framework based on voxel-points, which effectively improves computational efficiency and reduces computational parameters compared to traditional 3D convolution methods.

Compared with the method proposed by TSPNet [34] (our conference paper method), our approach fully utilizes attention mechanisms and combines scene and instance features to achieve accurate reconstruction of indoor scenes. Our proposed method adds object detection and object refinement modules. The object refinement module considers the inherent sparsity of voxel space and employs a spatial attention mechanism to focus on effective voxels. Compared with the TSPNet [34] method, the scene reconstruction instance results have more precise geometric details. For SSC mIoU, our network performance improved by 0.2% and 0.2% on the NYU dataset and NYUCAD dataset, respectively.

II. RELATED WORK

Currently, 3D semantic scene complementation methods can be summarized into three main types according to the different types of input data: depth map-based semantic scene completion methods, depth map joint color image-based semantic scene completion methods and point cloud-based semantic scene completion methods.

A. DEPTH MAP-BASED SEMANTIC SCENE COMPLETION

Originally pioneered by Song et al., SSCNet [8] takes a single-frame depth map as input and generates categories corresponding to all voxels in the camera view cone. The algorithm employs the Flipped Truncated Signed Distance Function (f-TSDF) to encode the voxels of the depth map, and uses the Extended Contextual Convolution Module to simultaneously perform voxel mesh occupancy and semantic labelling prediction of the scene. VVNet [9] directly performs two-dimensional convolution of the depth map to obtain a two-dimensional feature, and projects the feature to the three-dimensional voxel space, which reduces the computational cost, deepens the depth of the feature extraction network, and makes the feature extraction results more adequate. ESSCNet [10] is an efficient semantic scene complementation model based on group convolution, which replaces the traditional 3D convolution with spatial group convolution, divides the input voxels into different groups, and uses sparse convolution for each group to perform feature extraction separately, thus effectively improving the computational efficiency while guaranteeing the computational accuracy. CCPNet [4] is a cascaded context pyramid network, the algorithm not only improves the labelling consistency in the pyramid context, but also proposes the Guided Residual Refinement (GRR) module to incrementally recover the fine-grained structure of the scene, and achieves competitive results on the SUNCG and NYUv2 datasets, with particular advantages in the generation of scene details. ForkNet [11] is based on a single encoder as well as three juxtaposed decoders. The three decoder branches predict incomplete surface geometries, geometric volumes, and semantic volumes, respectively, and multiple discriminators are introduced to improve the accuracy and realism of the semantic scene-completion task. In PALNet [13], Li et al. designed a new loss function PA-Loss based on the semantic relationship between voxels and surrounding voxels, which adaptively adjusts the weight of the voxel in the cross-entropy loss through the Local Geometric Anisotropy (LGA) of the voxel, so that the network focuses on the voxels located at the junction of the objects, which is conducive to the recovery of the object's boundary information and the information of the scene corners.

B. DEPTH MAP JOINT COLOR IMAGE-BASED SEMANTIC SCENE COMPLETION

Combining RGB images with depth maps can improve the network's performance in recognizing surface texture and color features. TS3D [14] is based on two-stream convolution, which maps the RGB image semantic segmentation results onto a 3D mesh generated from a depth map to obtain incomplete semantic corpora. The complete semantic scene information is then inferred using a context-aware 3DCNN. Experiments show that introducing RGB images as inputs can significantly improve the SSC task by 9.4% compared to the 2nd place on the NYUv2 dataset. DDRNet [15] uses Dimensional Decomposition Residual (DDR) to replace the standard three-dimensional convolution operation. By splitting the standard 3D convolution kernel into three 1D convolution kernels in series, the DDR module drastically reduces the convolution parameters and lowers the computational consumption while keeping the sensory field unchanged. Based on DDRNet, Li et al. [16] combined the DDR module with the attention mechanism to design a new basic unit of feature extraction, the AIC (Anisotropic convolution) module.The AIC module adaptively assigns weights to the three convolution kernels of the split, which further improves the accuracy of the semantic complementation. GRFNet [17] s the first to use Gated Recurrent Unit (GRU), which is extended based on the DDRNet network, to improve the multiscale fusion strategy and construct a multimodal feature fusion module with autonomous selection and adaptive memory preservation. In addition, different levels of features are fused by introducing non-significance parameters and further propose a multi-stage fusion strategy. SISNet [18] is an iterative semantic complementation network for scene-to-instance and instance-to-scene. Li [19] proposed AMFNet, a multimodal fusion network based on the attention mechanism. the algorithm uses 2D segmentation results to guide the SSC task. OccDepth [20] utilizes implicit depth information in binocular images to reconstruct 3D geometric structures. The stereo soft feature alignment module (Stereo-SFA) better fuses 3D depth-aware features by learning the correspondence between binocular images. VoxFormer [22] first employs deep estimation network to obtain Query Proposals for the visible region, and then applies a masked autoencoder to complete the complementation by propagating the information to all voxels through a self-attentive mechanism.

C. POINT CLOUD-BASED SEMANTIC SCENE COMPLETION

S3CNet [23] employs a bird's eye view of efficient sparse 3D tensor projections obtained through point clouds for semantic segmentation, and the resulting 2D segmentation results are used to enhance 3D SSC. Zhong and Zeng [24] proposed a scene complementation network IPF-SPC-Net that fuses RGB image texture information with point cloud geometry information. Yan et al. [25] proposed JS3C-Net, a semantic segmentation framework for sparse radar point clouds with context shape prior. Rist et al. [26] proposed LMSC-Net, a semantic scene-completion network based on local depth implicit functions. The method uses a non-somatized continuous scene representation and introduces free spatial information as a supervisory signal, which yielded good experimental results on the outdoor scene dataset Semantic KITTI. SCPNet [27] enhances SSC from the aspects of the completion network redesign, dense-to-sparse knowledge distillation as well as completion label rectification.

CasFusionNet [28] is a novel cascaded network for point cloud semantic scene completion by dense feature fusion. A global completion module (GCM), a semantic

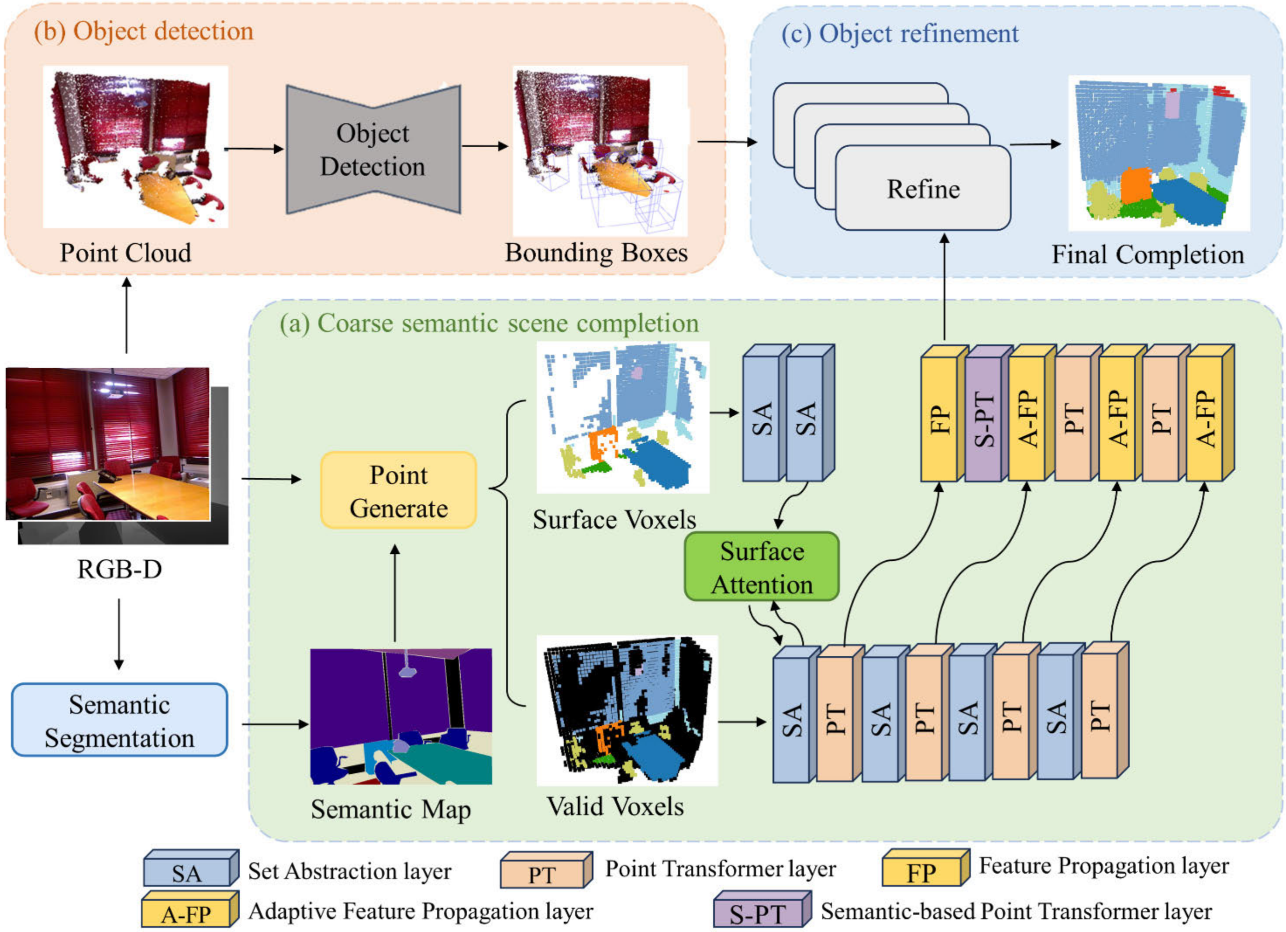


FIGURE 1. The architecture of ASPNet. ASPNet consists of three parts: (a) coarse semantic scene completion, where an SSC architecture based on voxel points and Transformer is employed; (b) object detection, where we output bounding boxes for each instance; (c) object refinement, where we perform corrections on voxels within the bounding boxes to produce the final result.

segmentation module (SSM) and a local refinement module (LRM) are designed and organized via dense feature fusion in each level, and cascade a total of four levels.

III. METHODS

The overall structure of ASPNet is shown in Figure 1. The entire network consists of two branches, namely the 3D object detection branch and the semantic scene completion branch based on voxel point representation. The results of these two branches are input into the geometric refinement module, which performs refinement and completion to obtain the final completion result. ASPNet consists of three parts: coarse semantic scene completion, object detection, and object refinement. For object detection, we output the bounding box of each instance. For coarse semantic scene completion, we use a lightweight SSC architecture based on voxel points and Transformer. Finally, we use the object refinement module to correct the voxels located inside the bounding box to generate the final result.

A. COARSE SEMANTIC SCENE COMPLETION

1) 2D SEMANTIC SEGMENTATION MODULE

We introduce RGB images as texture features to enhance the network's understanding of local observations. By projecting the 2D semantic segmentation results in Figure 2 onto the 3D voxel space, ASPNet can directly obtain the semantic labels corresponding to the visible surface voxels, thereby providing rich prior information for subsequent 3D convolutional networks.

Firstly, a pair of RGB-D images from a single perspective are used as input, and the SSC branch based on voxel point representation is used to complete 2D semantic segmentation of the RGB-D images. The output shape of the 2D semantic segmentation is $M_S \in R^{12\times480\times640}$, where 12 represents the preset number of categories in the scene. ASPNet chose the DeeplabV3+ [32] model based on Resnet-101 [31] as the semantic segmentation network. We replace the traditional two-dimensional convolutional layer in the network with the Shapeconv layer to achieve more accurate semantic segmentation. The Shapeconv [33] layer focuses more

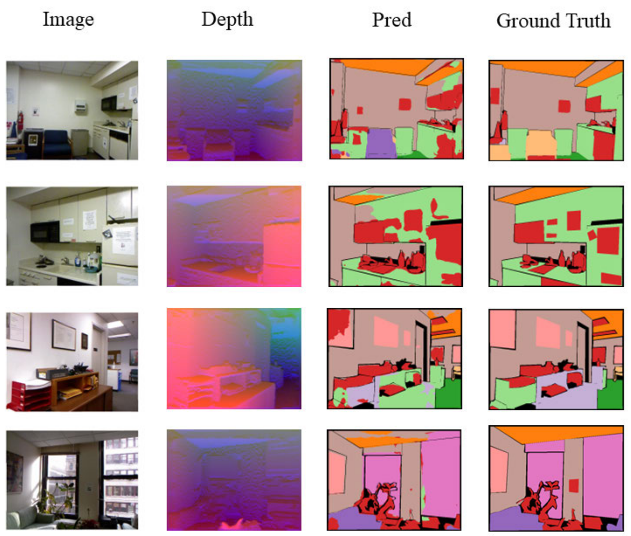


FIGURE 2. Results of semantic segmentation. From left to right are the image, depth, semantic segmentation results, and ground truth.

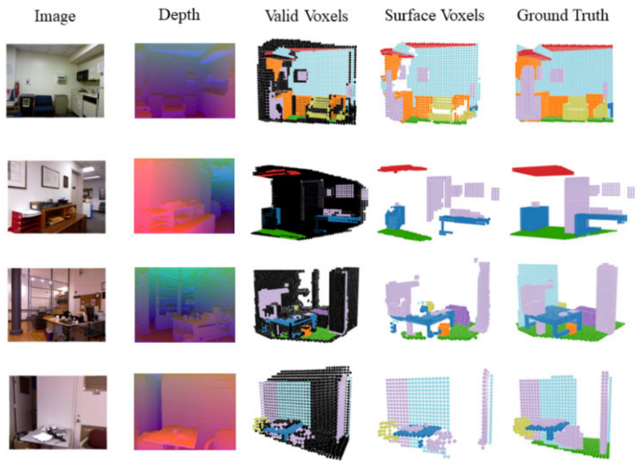


FIGURE 3. Voxel-points located on visible surfaces.

on shape information compared to traditional 2D convolutional layers, helping to reduce the possibility of extracting completely different features due to the distance of similar objects.

2) VOXEL-POINTS GENERATION

The voxel-based scene semantic reconstruction method usually converts the input depth map into TSDF format data. In SSC tasks, all voxels in TSDF do not have the same importance. Invalid voxels located between the camera and the observation surface that can be directly identified as air [34]. Therefore, ASPNet only retains voxels located on the observable surface and behind it, and converts the retained effective voxels into point clouds $P^{valid} \in R^{N_{valid} \times C_{voxel}}$ in Figure 3. For any ''voxel-points'' $p_i \in P^{valid}$, which contains the eigenvector of $c_i \in R^{17}$, specifically, c_i can be expressed as:

$$c_i = (x_i, y_i, z_i, t_i, h_i, s_i) \tag{1}$$

where the voxel v_i mapped by p_i comes from a 3D feature volume of size $60 \times 36 \times 60$. (x_i, y_i, z_i) is the normalized $x - y - z$ index of v_i in the feature volume. t_i is the TSDF value of v_i. h_i is the height value at which v_i is located.

In ASPNet, the normalized center of (x_i, y_i, z_i) is the center value of P^{valid}, while the normalized center of h_i is 36. h_i can serve as prior information to describe the position of objects in the scene, effectively distinguishing categories with significantly different height values, such as floors and ceilings. ASPNet introduces $s_i \in R^{12}$ as the semantic feature carried by each ''voxel-points''. For the ''voxel-points'' distributed on the visible surface, their s_i is the semantic feature vector corresponding to v_i in $V_S \in R^{12\times 60\times 36\times 60}$, while for the points distributed behind the visible surface, their s_i is a zero vector.

3) ADAPTIVE SURFACE-ATTENTION MODULE

Inspired by the Surface-Attention module [34] and ResNet [31] structure, we have designed an adaptive surface-attention module. This module adaptively transfers local surface features to internal points, thereby supplementing the features of internal ''voxel-points''. Specifically, the process of Adaptive Surface-Attention can be expressed as follows:

$$y_i = \sum\nolimits_{f_j \in \chi'(i)} \rho\left(\gamma(\varphi(q_i) - \omega(f_j)) + \delta'\right) \odot (\alpha(f_j) + \delta') \tag{2}$$

$$\delta' = \theta\left(f_j^{xyz} - q_i^{xyz}\right) + \varepsilon(d_{cos}(f_j, q_i)) \tag{3}$$

where y_i is the output vector, f_j is local feature, q_i^{xyz} is the 3D coordinate of the surface key points, φ, ω, α is the encoding operation for individual vectors, usually the MLP layer or linear layer, δ is the positional encoding, and ρ is the normalization function (usually the softmax function). ε is also an MLP operation, d_{cos} is the cosine distance corresponding to two local features.

In addition, $\chi'(i)$ is the union of the k (preset value) points closest to each other in the 3D and high-dimensional feature spaces of $F_{surafce}$ and q_i. $\chi'(i)$ can be represented as follows:

$$\chi'(i) = K^{near}(q_i^{xyz}, F_{surface}^{xyz}) \cup K^{near}(\tau(q_i), \tau(F_{surface})) \tag{4}$$

where K^{near} represents k nearest neighbor points, and q_i^{xyz} and F_{valid}^{xyz} represent the 3D coordinate values corresponding to q_i and $F_{surface}$. τ represents an MLP operation that maps local features q_i and $F_{surface}$ to the same space.

The structure of Adaptive Surface-Attention is shown in Figure 5. Through the attention mechanism, the surface attention module supplements voxel-points features with weak texture features, effectively improving the accuracy of scene semantic reconstruction.

4) SEMANTIC-BASED POINT TRANSFORMER MODULE

In order to suppress the impact of different types of key ''voxel-points'' with deeper network layers on the current voxel points, we replace the last layer of ASPNet's Point Transformer layer with a Semantic based Point Transformer

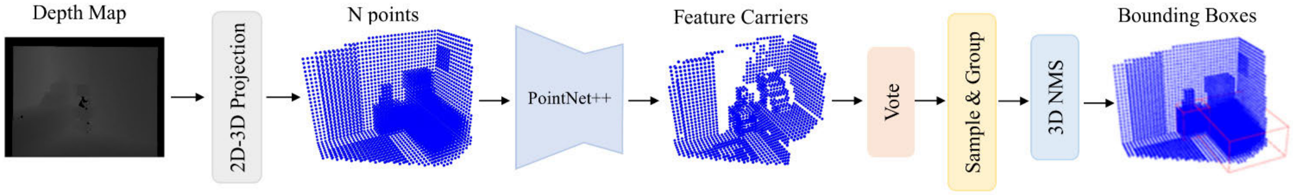


FIGURE 4. The architecture of object detection module in ASPNet.

(SPT) layer. In the SPT layer, the impact of non similar points is reduced by adding binary judgments. The SPT module can be represented as:

$$y_i = \sum\nolimits_{f_j \in \chi_s(i)} \partial \odot \left(\alpha\left(f_j\right) + \delta\right) + \sum\nolimits_{f_j \in \chi_d(i)} \partial \odot \delta \tag{5}$$

$$\partial = \left(\gamma\left(\varphi\left(f_i\right) - \omega\left(f_j\right)\right) + \delta\right) \tag{6}$$

$$\chi_s\left(i\right) = \{f_j | f_j \in \chi\left(i\right) \cap \eta\left(\varphi\left(f_i\right) - \omega\left(f_j\right)\right) \geq 0.5\} \tag{7}$$

$$\chi_d\left(i\right) = \{f_j | f_j \in \chi\left(i\right) \cap \eta\left(\varphi\left(f_i\right) - \omega\left(f_j\right)\right) < 0.5\} \tag{8}$$

where $\chi_s(i)$ represents the key "voxel-points" that are judged to be of the same class among the nearest k points for the current point q_i, and $\chi_d(i)$ represents the key "voxel-points" that are judged to be non of the same class among the nearest k points for the current point q_i. η represents an MLP operation consisting of two linear layers and a sigmoid activation function. The SPT layer can aggregate and adjust the local features of similar "voxel-points", thereby improving the similarity of similar local features in high-dimensional space and avoiding the occurrence of outliers and "noise".

B. OBJECT DETECTION MODULE

The object detection module in ASPNet is the same as Votenet [29]. Votenet is currently an effective algorithm in the field of 3D object detection, which can directly process raw point cloud data without relying on 2D detectors. As one of our branches focuses on object detection, Votenet's robustness in detecting 3D objects aligns well with our goal. The network first uses the PointNet++ network [35] to extract point cloud features, then uses the Hough voting mechanism to obtain voting points, clusters to obtain voting clusters, and finally extracts the bounding box and category for each cluster's feature prediction. The Hough voting mechanism efficiently identifies object centers, contributing to accurate bounding box predictions. Its ability to capture fine-grained details and maintain spatial accuracy is crucial for our task of semantic scene completion. Figure 4 shows the architecture of the object detection module. Figure 8 shows the results of the object bounding box obtained by the object detection module in point cloud and voxel space.

Our object detection module borrows VoteNet [29] pretrained models. Due to its good performance, these weights are frozen during the training of our network. It is worth noting that the object detection module is weakly coupled with the semantic completion branch, so the former can be replaced with any accurate point cloud object detection method without the need to retrain the network parameters of the latter. In addition, the object surface point cloud located in the bounding box will be extracted and used as supplementary information input for subsequent geometric refinement modules.

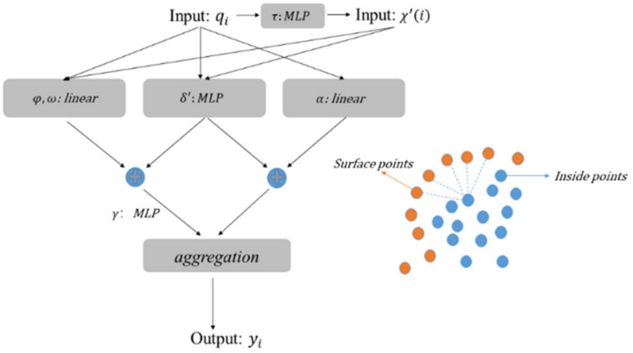


FIGURE 5. The architecture of adaptive surface-attention module.

C. OBJECT REFINEMENT MODULE

After completing the generation of S_{coarse}, the object refinement module based on instance level features is used to refine the voxels located within the bounding box. Using an "encoder-decoder", the object refinement module in ASPNet fully utilizes the features of S_{coarse} and surface point cloud P_{in}. Firstly, based on the object bounding box in voxel space, the voxels located within the box in S_{coarse} are intercepted, which can be represented as follows:

$$S_i = S_{coarse}\left[x_{tg}, y_{tg}, z_{tg}\right] \tag{9}$$

where $x_{tg}\epsilon\left[x^i_{min}, x^i_{max}\right]$, $y_{tg}\epsilon\left[y^i_{min}, y^i_{max}\right]$, $z_{tg}\epsilon\left[z^i_{min}, z^i_{max}\right]$, S_i represents the voxel to the i-th bounding box, and $(x^i_{min}, y^i_{min}, z^i_{min})$, $(x^i_{max}, y^i_{max}, z^i_{max})$ represents the minimum and maximum coordinates of the corresponding bounding box in voxel space, respectively.

Due to the roughness of $S_i \in R^{12\times H_i \times W_i \times L_i}$, we have introduced a surface local point cloud P located within the detection box as another input to this module. This process can be represented by the following functions:

$$S_i^{''} = G\left(f\left(S_i\right), h\left(P_i\right)\right) + S_i \tag{10}$$

where $S_i^{''}$ is the voxel result refined by the object refinement module, f, h are the encoding operations for voxels and point

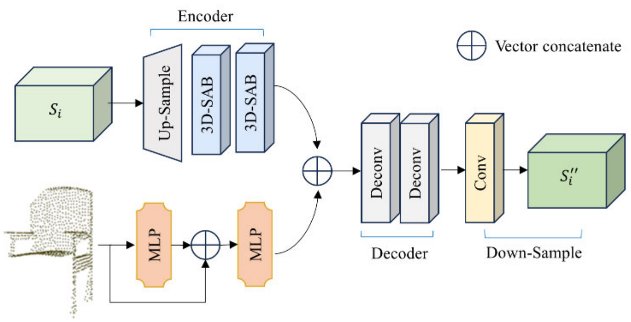


FIGURE 6. The architecture of object refinement module.

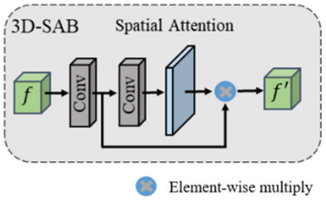


FIGURE 7. The architecture of 3D-SAB module.

cloud inputs, and G is the decoding operation for the obtained high-dimensional features.

Due to the sparsity in voxel representation methods, we interpolate and upsample S_i to obtain $S_i^2 \in R^{12\times 4H_i\times 4W_i\times 4L_i}$. Meanwhile, inspired by CBAM [36], we designed a 3D Spatial Attention Block (3D-SAB) for extracting features from S_i^2, whose structure is shown in Figure 7. By utilizing spatial attention mechanism, the network can adaptively focus on voxels that contribute significantly to the bounding box.

D. TRAINING LOSS

The training process of ASPNet consists of three parts: a 2D semantic segmentation network, a 3D object detection network, and coarse semantic scene completion network. The network training of these three parts is independent of each other.

1) 2D SEMANTIC SEGMENTATION NETWORK

ASPNet uses the cross entropy loss between the output image and the real semantic labels. Specifically, the loss function of the 2D semantic segmentation module can be expressed as:

$$L_{2D} = l_{ce}(M_S, M_S^{gt}) \tag{11}$$

where l_{ce} is the cross-entropy loss function.

2) OBJECT DETECTION NETWORK

The loss function used is the same as VoteNet [29], which can be expressed as:

$$L_{stage-1} = \lambda_1 L_{box} + \lambda_2 L_{sem-cls} + \lambda_3 L_{obj-cls} + L_{vote-reg} \tag{12}$$

where $L_{vote-reg}$ represents the voting loss, $L_{obj-cls}$ is the binary cross entropy loss used to determine whether there are objects in the current candidate box, and $L_{sem-cls}$ is also the cross entropy loss used to determine the category of objects in the current candidate box. L_{box} is composed of multiple L1 losses. λ_1, λ_2, λ_3 are the weights, with values of 0.5, 1, and 0.1 in ASPNet.

3) COARSE SEMANTIC SCENE COMPLETION NETWORK

The loss function is the sum of SSC loss function L_{SSC} and SPT loss function L_{SPT}:

$$L = L_{SSC} + L_{SPT} \tag{13}$$

where L_{SSC} can be represented as:

$$L_{SSC} = \frac{1}{N_{valid}}\left(\sum\nolimits_{i,j,k} m_{i,j,k} * l_{ce}(s_{i,j,k}, s^{gt}_{i,j,k})\right) \tag{14}$$

where N_{valid} represents the number of effective voxel-points, $s_{i,j,k}$ represents the value of the predicted semantic scene completion result at (i, j, k), and $s^{gt}_{i,j,k}$ represents the ground truth. If (i, j, k) is an invalid voxel, the value of m is 0, otherwise the value of m is 1. L_{SPT} is used to supervise the binary classification network in the SPT module:

$$L_{SPT} = \frac{1}{N_{pairs}} \sum\nolimits_{0\leq i\leq \mathbb{N}^l} \sum\nolimits_{j\epsilon \chi_s(i)} l_{ce}(w_{i,j}, g_{i,j}) \tag{15}$$

where N_{pairs} represents the number of voxel points pairs involved, N^l represents the number of key points corresponding to the SPT layer, $w_{i,j}$ represents the binary prediction value of (i, j) point pairs, and $g_{i,j}$ is the ground truth.

IV. RESULTS AND DISCUSSION

A. DATASETS AND EVALUATION METRICS

We evaluated the proposed ASPNet on the NYU [37] and NYUCAD [38] datasets. The NYU dataset comprises 1449 scenes, divided into 795 for training and 654 for testing, following the same partitioning as SSCNet [8]. To mitigate measurement errors in the NYU dataset, NYUCAD employs 3D annotations to generate synthetic depth maps. Similar to SSCNet [8], our evaluation focuses on Scene Completion (SC) and Semantic Scene Completion (SSC). SC consolidates non-empty voxels into a category, assessing metrics such as IoU, recall, and precision. SSC evaluation computes IoU for each semantic category within the valid frustum, yielding the mean IoU (mIoU) through category averages. In addition, since the semantic scene completion method based on voxel points has already eliminated invalid voxels, ASPNet only needs to perform quantitative comparative evaluation on SSC tasks to fully demonstrate the accuracy of ASPNet in semantic scene completion.

B. IMPLEMENTATION DETAILS

In the experiment, we used RGB-D images with a resolution of 480 × 640 as input. Our model was implemented using Pytorch, and the entire training process was divided into three stages: training of the 2D semantic segmentation module and

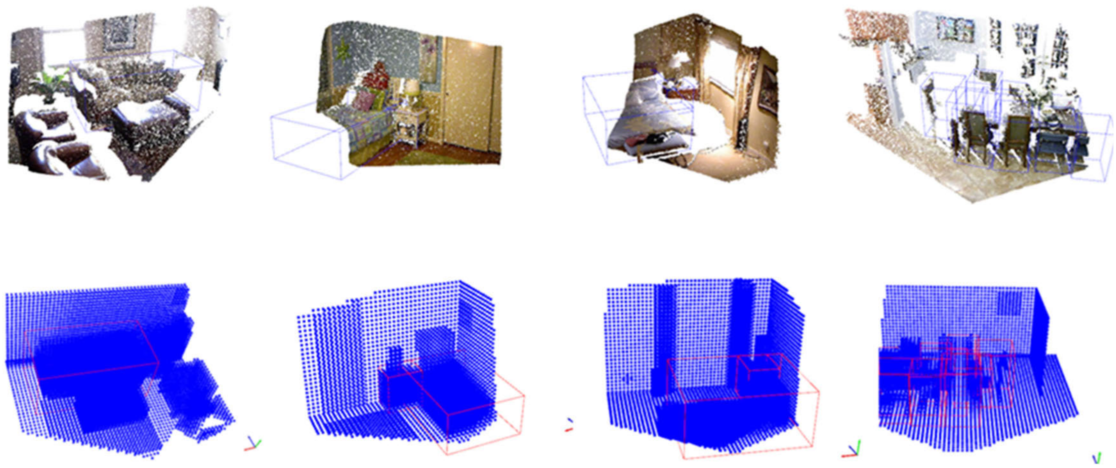


FIGURE 8. Qualitative results of the object detection module in ASPNet.

the object detection module, followed by the training of the semantic scene completion module. For the former, we used the Adam optimizer, a batch size of 4, an initial learning rate of 0.001, and a total of 200 training epochs. Similarly, for the latter, we employed the Adam optimizer with an initial learning rate of 0.001, a batch size of 4, and a total of 250 training epochs. The training was completed on an Ubuntu 18.04 system equipped with a single NVIDIA RTX 2080TI GPU. During the training of ASPNet, we performed data augmentation on the 3D input. For the set of ''voxel points,'' we rotated P^{valid} and $P^{surface}$ along the z-axis at 90° intervals, thereby expanding the training set. After completing the training, we evaluated the computational requirements and memory consumption of ASPNet.

C. DETAILS ON OBJECT DETECTION MODULE

Our object detection module borrows VoteNet [29] and fine-tunes its parameters to achieve the best results. The layer parameters of the Set Abstraction (SA) layer and Feature Propagation (FP) layer in the backbone network are the same as those in VoteNet. In the proposal step, we select 256 voting clusters to generate 256 proposals from the votes. We use a cluster sampling strategy of farthest point sampling (FPS) on seeds. We use a voting factor of 1, where the voting module generates one vote for each seed. We use a 1% Z-value of all points in the scan as the height feature. These parameter settings can achieve the best detection results (mAP 50.2%). Figure 8 shows the qualitative results of our object detection module. It is worth noting that we first pre-train the VoteNet network, and then during the overall network training process, the network parameters in VoteNet are frozen.

D. COMPARISONS WITH STATE-OF-THE-ART METHODS

We compared ASPNet and SOTA (state of the art) methods on the NYU dataset and NYUCAD dataset, and the results are shown in Tables 1 and 2. On the NYU dataset, the ASPNet method achieved SOTA (state of the art) effect in non iterative methods (One off). Specifically, compared to SketchNet [42], ASPNet exceeded 10.0% on SSC mIoU in the NYU dataset. Our network performs better in completing objects with smaller geometric dimensions (such as chairs, tables, etc.) than other networks. In addition, mainstream semantic scene completion algorithms based on 3D convolution have high computational overhead and long training time, while our network backbone is implemented based on one-dimensional convolution, so the computational cost is relatively low. To demonstrate this, we conducted a comparative evaluation of the computational cost (FLOPs) of ASPNet, as shown in Tables 3. Compared to SSCNet [8] based on 3D convolution, SketchNet [42], and the iterative SOTA algorithm SISNet DeepLabv3 [18], ASPNet significantly reduces computational cost and efficiently completes scene semantic reconstruction tasks without losing accuracy.

E. ABLATION STUDIES

We compared ASPNet and SOTA In order to prove the effectiveness of each module in ASPNet, we conducted ablation studies on the key modules of the network and evaluated the effectiveness of the Surface-Attention module, the SPT module, and the Adaptive Feature Propagation layer (A-FP).

We removed the Surface-Attention module from ASPNet and quantitatively compared it with ASPNet with the Surface-Attention module applied, and the results are shown in Table 4. As can be seen from the Table 4, the surface attention module improves the performance of SSC mIoU (for SSC mIoU: 1.0% increase on the NYU dataset). Therefore, the effectiveness of the surface attention module in ASPNet is proved by quantitative comparison. In order to demonstrate the effectiveness of the surface attention module, we replaced the SPT module in ASPNet with the regular Point Transformer layer and conducted a quantitative comparison with ASPNet with SPT applied. As can be seen from the Table 4, introducing semantic information into the Point Transformer layer can effectively improve the performance of semantic

TABLE 1. **SSC results on NYU dataset.**

Method (One-off)	ceil	floor	wall	win.	chair	bed	sofa	table	tvs	furn	objs.	avg.
SSCNet [8]	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 [10]	17.5	75.4	25.8	6.7	15.3	53.8	42.4	11.2	0	33.4	11.8	26.7
DDRNet [15]	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]	19.3	**94.8**	28.0	12.2	19.6	57.0	50.5	17.6	11.9	35.6	15.3	32.9
GRFNet [17]	24.0	91.7	33.3	19.0	18.1	51.9	45.5	13.4	13.3	37.3	15.0	32.9
AMFNet [19]	16.7	89.2	27.3	19.2	20.2	56.1	50.4	15.1	13.5	36.8	18.0	33.0
AICNet [16]	23.2	90.8	32.3	14.8	18.2	51.1	44.8	15.2	22.4	38.3	15.7	33.3
TS3D [14]	9.7	93.4	25.5	21.0	17.4	55.9	49.2	17.0	27.5	39.4	19.3	34.1
PALNet [13]	23.5	92.0	33.0	11.6	20.1	53.9	48.1	16.2	24.2	37.8	14.7	34.1
SATNet [41]	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 [11]	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 [4]	23.5	96.3	35.7	20.2	25.8	61.4	56.1	18.1	28.1	37.8	20.1	38.5
SketchNet [42]	43.1	93.6	40.5	24.3	30.0	57.1	49.3	29.2	14.3	42.5	28.6	41.1
IMENet [6]	43.6	93.6	42.9	31.3	36.6	57.6	48.4	32.1	16.0	47.8	**36.7**	44.2
PVA-Net [5]	51.4	94.0	49.9	15.9	41.9	68.3	58.8	35.4	12.9	48.5	29.1	46.0
CleanerS [43]	48.2	94.0	43.2	33.7	38.5	62.2	54.8	33.7	**39.2**	45.7	33.8	47.7
PCANet [44]	44.3	94.5	**50.1**	30.7	41.8	68.5	56.4	32.6	29.9	53.6	35.4	48.9
ASPNet (Ours)	**54.7**	93.7	47.0	**37.2**	**43.5**	**69.4**	**61.4**	**39.2**	29.3	**55.1**	32.3	**51.2**
Method (Iterative)												
SISNet-BiSeNet [18]	53.9	93.2	51.3	38.0	38.7	65.0	56.3	37.8	25.9	51.3	36.0	49.8
SISNet-DeepLabv3 [18]	54.7	93.8	53.2	41.9	43.6	66.2	61.4	38.1	29.8	53.9	40.3	52.4

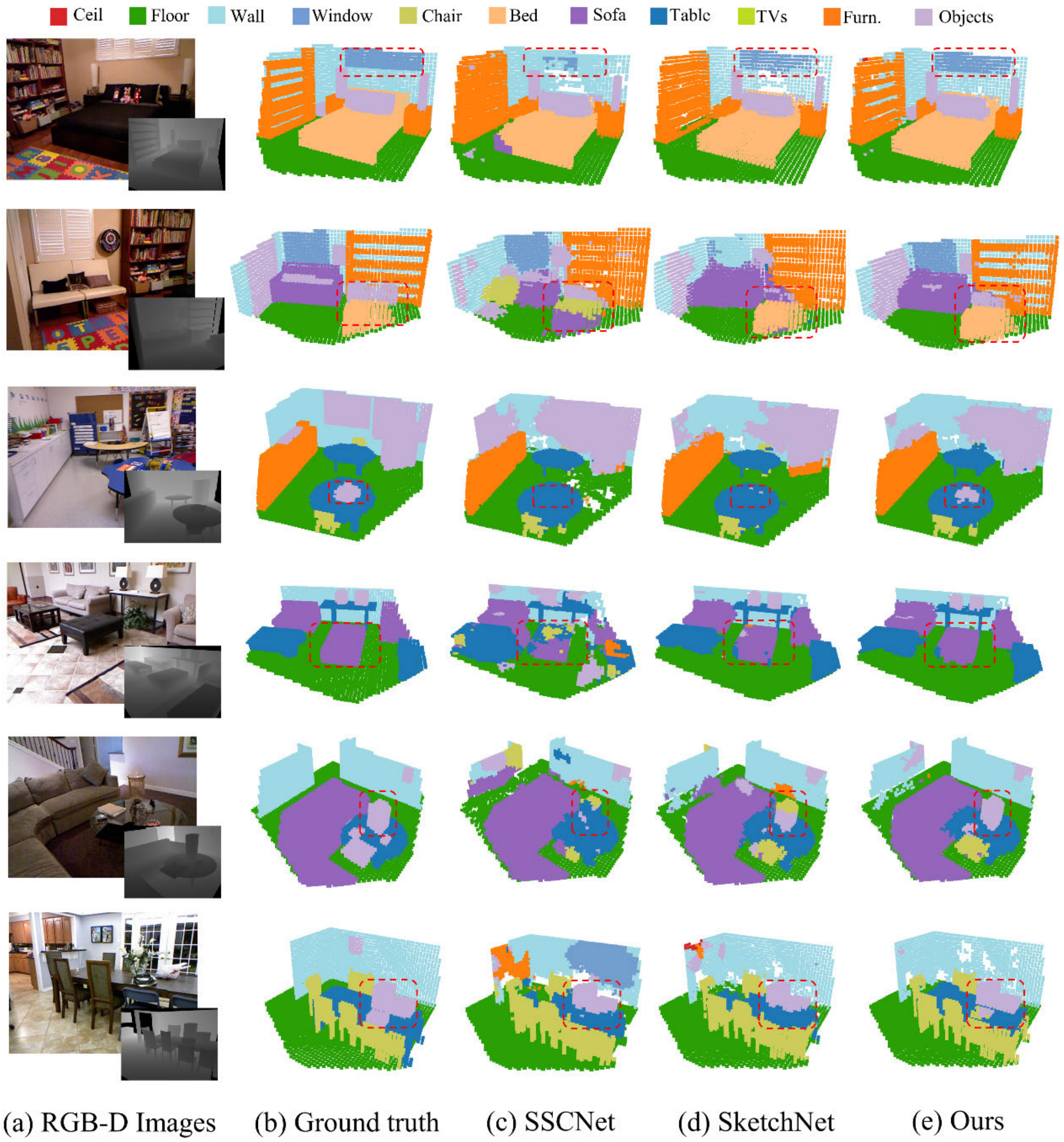
TABLE 2. **SSC results on NYUCAD dataset.**

Method (One-off)	ceil	floor	wall	win.	chair	bed	sofa	table	tvs	furn	objs.	avg.
SSCNet [8]	32.5	92.6	40.2	8.9	33.9	57.0	59.5	28.3	8.1	44.8	25.1	40.0
DDRNet [15]	54.1	91.5	56.4	14.9	37.0	55.7	51.0	28.8	9.2	44.1	27.8	42.8
GRFNet [17]	50.3	91.8	58.1	18.4	42.7	60.6	52.8	34.6	11.5	46.6	30.8	45.3
AICNet [16]	53.0	91.2	57.2	20.2	44.6	58.4	56.2	36.2	9.7	47.1	30.4	45.8
TS3D [14]	25.9	93.8	48.9	33.4	31.2	66.1	56.4	31.6	**38.5**	51.4	30.8	46.2
PALNet [13]	54.8	92.8	60.3	15.3	43.1	60.7	59.9	37.6	8.1	48.6	31.7	46.6
CCPNet [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
SketchNet [42]	59.7	94.3	64.3	32.6	**51.7**	72.0	68.7	45.9	19.0	**60.5**	**38.5**	55.2
PVA-Net [5]	**71.5**	94.1	**66.6**	23.7	60.0	**78.5**	**72.2**	**45.3**	16.7	60.1	36.9	**56.9**
ASPNet (Ours)	62.2	**95.4**	59.8	**41.4**	48.1	67.8	59.9	40.9	37.0	58.8	37.9	55.4
Method (Iterative)												
SISNet-BiSeNet [18]	65.6	94.4	67.1	45.2	57.2	75.5	66.4	50.9	31.1	62.5	42.9	59.9
SISNet-DeepLabv3 [18]	63.4	94.4	67.2	52.4	59.2	77.9	77.1	51.8	46.2	65.8	48.8	63.5

scene complementation (for SSC mIoU: 0.8% increase on the NYU dataset). To demonstrate the effectiveness of the adaptive feature propagation layer (A-FP), we replaced the A-FP layer in ASPNet with a regular FP layer and performed a quantitative comparison with ASPNet with the A-FP layer applied. As can be seen from the Table 4, the adaptive feature propagation layer improves the performance of SSC mIoU (for SSC mIoU: 0.5% increase on the NYU dataset.

We designed ablation studies on object refinement and surface point clouds in refining. In the object refinement module, we sum the output results element-by-element with the rough results located in the corresponding detection boxes. As can be seen from Table 5, the element-by-element addition-based geometric refinement strategy improves the performance of SSC mIoU compared to the Semantic Scene Completion (SSC) results without the object refinement module (for SSC mIoU: an increase of 4.2% and 5.2% on the NYU dataset and the NYUCAD dataset by 4.2% and 5.2%, respectively). In addition, the surface point cloud of an object can provide correction information and complement geometric details to the relatively low resolution rough complementary results. To demonstrate this, we compare the effect of semantic complementation before and after adding point cloud branches. As shown in Tables 5, the introduction of surface point clouds by the object refinement module can effectively improve the semantic complementation effect.

For SSC mIoU: 0.4% improvement on the NYU dataset and 1.0% improvement on the NYUCAD dataset.



(a) RGB-D Images (b) Ground truth (c) SSCNet (d) SketchNet (e) Ours

FIGURE 9. Qualitative results on the NYU dataset.

TABLE 3. Calculation efficiency of different methods.

Methods	FLOPs	SSC mIoU
SSCNet [8]	163.8G	24.7
SketchNet [42]	293.7G	41.1
SISNet-DeepLabv3 [18]	733.6G	52.4
ASPNet (Ours)	27.9G	51.1

TABLE 4. Ablation study on key modules.

Methods	DATA	SSC mIoU
ASPNet	NYU	51.1
ASPNet(w/o SA)	NYU	50.1
ASPNet(w/o SPT)	NYU	50.3
ASPNet(w/o A-FP)	NYU	50.6

F. QUALITATIVE VISUALIZATION

As can be seen from some of the visualization results in Figure 9, we qualitatively evaluate the effectiveness of ASPNet. Generally, we can see that ASPNet complements and completes the geometric details of small objects, such as ‘‘chair’’ and ‘‘table’’ in rows 3, 5 and 6, compared to SSCNet. In different scenarios, the semantic complementation results of our network for different objects are richer in details, more delicate and complete. Not only that, the method in this paper performs better than SSCNet and SketchNet on objects such as walls, windows, chairs, beds, etc., but also when there are more furniture in the scene, the generated voxels are more

TABLE 5. Ablation study on object refinement and point cloud in refining.

Methods	DATA	SSC mIoU
ASPNet	NYU	51.1
ASPNet (w/o Refinement)	NYU	42.8
ASPNet (w/o Point cloud)	NYU	42.4
ASPNet	NYUCAD	55.4

concise and complete in terms of object shape completion and semantic segmentation.

G. QUALITATIVE ANALYSIS

ASPNet can recognize and reconstruct physically smaller objects in the scene, reconstructing clearer 3D shape boundaries of small objects, such as the sofa in row 4 and the object in row 5. We attribute this to the 3D object detection branch that can provide boundary constraints and object refinement module. ASPNet excels in handling low texture regions and small objects with color features, which is attributed to our method utilizing 2D semantic segmentation results to provide guidance information. Due to the fact that RGB images carry more details, such as color and texture, this is beneficial for semantic information, as can be seen from the results of bed in row 2 and sofa in row 4 categories.

The surface attention module utilizes feature rich surface voxel points to supplement internal voxel points, which helps identify different categories, such as chairs and sofas, even if their colors and textures are very similar. Our method can handle both large object windows in row 1 and small object windows in row 3 very well, especially small object windows.

V. CONCLUSION

In this paper, we propose an efficient Attention-based Semantic Point Completion Network (ASPNet). With the object detection and instance-level refinement modules, our network is able to recover more geometric details and generate sharper object boundaries. Considering the redundancy of matrix-based representation, ASPNet eliminates the invalid voxels located in front of the visible surface and converts the remaining valid voxels into point cloud data. Based on this operation, ASPNet updates the 3D convolution-based feature extraction main frame to a 1D convolution-based main frame, which effectively improves the computational efficiency. Considering the difference between "voxel points" and traditional surface point clouds in the SSC task, ASPNet utilizes the surface attention module to supplement the internal "voxel points" with semantically rich surface "voxel-points". "The SPT module achieves feature aggregation of neighboring keypoints of the same category by introducing semantic information into the traditional Point Transformer layer. In addition, considering the defective type of the interpolation strategy of the traditional Feature Propagation Layer (FP Layer), ASPNet employs an attention-based interpolation algorithm to realize Adaptive Feature Propagation (A-FP). The quantitative results on NYU as well as NYUCAD datasets also demonstrate that ASPNet not only achieves SOTA reconstruction in non-iterative SSC algorithms, but also far outperforms 3D convolution-based SSC algorithms in terms of computational efficiency.

REFERENCES

[1] R. Fu, H. Wu, M. Hao, and Y. Miao, "Semantic scene completion through multi-level feature fusion," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Kyoto, Japan, Oct. 2022, pp. 8399–8406.

[2] J. Li, P. Wang, K. Han, and Y. Liu, "Anisotropic convolutional neural networks for RGB-D based semantic scene completion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8125–8138, Nov. 2022.

[3] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 77–85.

[4] P. Zhang, W. Liu, Y. Lei, H. Lu, and X. Yang, "Cascaded context pyramid for full-resolution 3D semantic scene completion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 7800–7809.

[5] J. Tang, X. Chen, J. Wang, and G. Zeng, "Not all voxels are equal: Semantic scene completion from the point-voxel perspective," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2352–2360.

[6] J. Li, L. Ding, and R. Huang, "IMENet: Joint 3D semantic scene completion and 2D semantic segmentation through iterative mutual enhancement," in *Proc. Thirtieth Int. Joint Conf. Artif. Intell.*, Montreal, QC, Canada, Aug. 2021, pp. 793–799.

[7] J. Hou, A. Dai, and M. Nießner, "RevealNet: Seeing behind objects in RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 2095–2104.

[8] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 190–198.

[9] Y. Guo and X. Tong, "View-volume network for semantic scene completion from a single depth image," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 726–732.

[10] J. Zhang, H. Zhao, A. Yao, Y. Chen, L. Zhang, and H. Liao, "Efficient semantic scene completion network with spatial group convolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 749–765.

[11] Y. Wang, D. J. Tan, N. Navab, and F. Tombari, "ForkNet: Multi-branch volumetric semantic completion from a single depth image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Oct. 2019, pp. 8607–8616.

[12] X. Chen, Y. Xing, and G. Zeng, "Real-time semantic scene completion via feature aggregation and conditioned prediction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Abu Dhabi, United Arab Emirates, Oct. 2020, pp. 2830–2834.

[13] J. Li, Y. Liu, X. Yuan, C. Zhao, R. Siegwart, I. Reid, and C. Cadena, "Depth based semantic scene completion with position importance aware loss," *IEEE Robot. Autom. Lett.*, vol. 5, no. 1, pp. 219–226, Jan. 2020.

[14] M. Garbade, Y.-T. Chen, J. Sawatzky, and J. Gall, "Two stream 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 416–425.

[15] J. Li, Y. Liu, D. Gong, Q. Shi, X. Yuan, C. Zhao, and I. Reid, "RGBD based dimensional decomposition residual network for 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7685–7694.

[16] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic convolutional networks for 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 3348–3356.

[17] Y. Liu, J. Li, Q. Yan, X. Yuan, C. Zhao, I. Reid, and C. Cadena, "3D gated recurrent fusion for semantic scene completion," 2020, *arXiv:2002.07269*.

[18] Y. Cai, X. Chen, C. Zhang, K.-Y. Lin, X. Wang, and H. Li, "Semantic scene completion via integrating instances and scene in-the-Loop," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 324–333.

[19] S. Li, C. Zou, Y. Li, X. Zhao, and Y. Gao, "Attention-based multi-modal fusion network for semantic scene completion," in *Proc. AAAI Conf. Artif. Intell.*, New York, NY, USA, 2020, pp. 11402–11409.

[20] R. Miao, W. Liu, M. Chen, Z. Gong, W. Xu, C. Hu, and S. Zhou, "OccDepth: A depth-aware method for 3D semantic scene completion," 2023, *arXiv:2302.13540*.

[21] A.-Q. Cao and R. de Charette, "MonoScene: Monocular 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 3981–3991.

[22] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 9087–9098.

[23] R. Cheng, C. Agia, Y. Ren, X. Li, and L. Bingbing, "S3CNet: A sparse semantic scene completion network for LiDAR point clouds," in *Proc. Conf. Robot Learn.*, Cambridge, MA, USA, 2020, pp. 2148–2161.

[24] M. Zhong and G. Zeng, "Semantic point completion network for 3D semantic scene completion," in *Proc. Eur. Conf. Artif. Intell.*, Santiago de Compostela, Spain, Aug. 2020, pp. 2824–2831.

[25] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep LiDAR point cloud segmentation via learning contextual shape priors from scene completion," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, pp. 3101–3109.

[26] C. B. Rist, D. Emmerichs, M. Enzweiler, and D. M. Gavrila, "Semantic scene completion using local deep implicit functions on LiDAR data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7205–7218, Oct. 2022.

[27] Z. Xia, Y. Liu, X. Li, X. Zhu, Y. Ma, Y. Li, Y. Hou, and Y. Qiao, "SCPNet: Semantic scene completion on point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 17642–17651.

[28] J. Xu, X. Li, Y. Tang, Q. Yu, Y. Hao, L. Hu, and M. Chen, "CasFusionNet: A cascaded network for point cloud semantic scene completion by dense feature fusion," in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, 2023, pp. 3018–3026.

[29] C. R. Qi, O. Litany, K. He, and L. Guibas, "Deep Hough voting for 3D object detection in point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Oct. 2019, pp. 9276–9285.

[30] Z. Zhang, B. Sun, H. Yang, and Q. Huang, "H3DNet: 3D object detection using hybrid geometric primitives," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., Aug. 2020, pp. 311–329.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 833–851.

[33] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 7068–7077.

[34] R. Fu, H. Wu, M. Hao, and Y. Miao, "Semantic scene completion with point cloud representation and transformer-based feature fusion," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Kuala Lumpur, Malaysia, Oct. 2023, pp. 3369–3373.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," 2017, *arXiv:1706.02413*.

[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.

[37] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Florence, Italy, Oct. 2012, pp. 746–760.

[38] M. Firman, O. M. Aodha, S. Julier, and G. J. Brostow, "Structured prediction of unobserved voxels from a single depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 5431–5440.

[39] A. Dourado, H. Kim, T. E. de Campos, and A. Hilton, "Semantic scene completion from a single 360-degree image and depth map," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, Valletta, Malta, 2020, pp. 36–46.

[40] J. Gong, J. Xu, X. Tan, J. Zhou, Y. Qu, Y. Xie, and L. Ma, "Boundary-aware geometric encoding for semantic segmentation of point clouds," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, pp. 1424–1432.

[41] S. Liu, Y. Hu, Y. Zeng, Q. Tang, B. Jin, Y. Han, and X. Li, "See and think: Disentangling semantic scene completion," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2018, pp. 261–272.

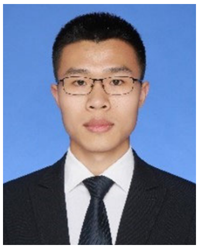
[42] X. Chen, K.-Y. Lin, C. Qian, G. Zeng, and H. Li, "3D sketch-aware semantic scene completion via semi-supervised structure prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 4192–4201.

[43] F. Wang, D. Zhang, H. Zhang, J. Tang, and Q. Sun, "Semantic scene completion with cleaner self," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 867–877.

[44] J. Li, Q. Song, X. Yan, Y. Chen, and R. Huang, "From front to rear: 3D semantic scene completion through planar convolution and attention-based network," *IEEE Trans. Multimedia*, vol. 25, pp. 8294–8307, 2023.

[45] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 567–576.

YUBIN MIAO received the Ph.D. degree in mechanical design and automation from Dalian University of Technology, in July 2001. From 2001 to 2003, he was a Postdoctoral Researcher with the Institute of Robotics, School of Mechanical Engineering, Shanghai Jiao Tong University, where he is currently with the School of Mechanical Engineering. His main research interests include the field of 3D point cloud analysis and processing and semantic scene completion.



JUNKANG WAN was born in Hubei, China, in 2000. He received the B.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2022, where he is currently pursuing the M.S. degree in mechanical engineering. His research interest includes 3D point cloud completion.

JUNJIE LUO received the B.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2022, where he is currently pursuing the M.S. degree in mechanical engineering. His research interests include the semantic segmentation of 2D images and 3D semantic scene completion based on deep leaning.

HANG WU received the B.S. and M.S. degrees from Shanghai Jiao Tong University, in 2016 and 2019, respectively, where he is currently pursuing the Ph.D. degree in mechanical engineering. His research interests include 3D vision, robotics, and deep learning.

RUOCHONG FU received the B.S. and M.S. degrees in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China. His research interests include semantic scene completion and 3D point cloud completion.