Point Cloud Semantic Scene Completion with Prototype-Guided Transformer

Chenghao Fang[1], **Jianqing Liang**[1*], **Jiye Liang**[1], **Zijin Du**[3], **Feilong Cao**[2]
[1]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi Taihang Laboratory, School of Computer and Information Technology, Shanxi University
[2]School of Mathematical Sciences, Zhejiang Normal University
[3]School of Engineering Science, University of Chinese Academy of Sciences
chenghaofang1014@163.com, liangjq@sxu.edu.cn, ljy@sxu.edu.cn, thetempest0302@163.com, icteam@163.com

Abstract

Semantic scene completion simultaneously reconstructs the shapes of missing regions and predicts semantic labels for the entire 3D scene. Although point cloud-based methods are more efficient than voxel-based methods, existing point cloud-based approaches largely fail to fully leverage semantic information. To address this challenge, we propose a **Proto**type-Guided Trans**former** (**ProtoFormer**) that encodes semantic information into a set of semantic prototypes to guide the underlying Transformer for semantic scene completion. Specifically, we leverage semantic prototypes to enhance information from both geometric and semantic perspectives, and integrate the top-K attention mechanisms to guide scene completion and semantic awareness. Extensive qualitative and quantitative experimental results demonstrate that ProtoFormer outperforms state-of-the-art approaches with low complexity.

Code — https://github.com/doldolOuO/ProtoFormer

Introduction

Semantic scene completion (SSC) aims to recover a complete scene from an incomplete one while simultaneously understanding the semantics of the completed result. This is crucial for environmental perception and intelligent decision-making in fields such as autonomous driving (Geiger, Lenz, and Urtasun 2012; Chen et al. 2024), robot navigation (Varley et al. 2017; Liang, Chen, and Song 2021; Cheng et al. 2022), and ancient architectural restoration (Li et al. 2025). With the rapid development of 3D sensors (such as RGB-D cameras and LiDAR) and deep learning techniques, SSC has emerged as a core task in 3D scene understanding (Liang et al. 2023; Du et al. 2024) and has attracted extensive attention from researchers.

Early SSC methods predominantly employed voxel-based representations, utilizing 3D convolutional networks to predict spatial occupancy and semantics. SSCNet (Song et al. 2017) is the first voxel-based method for semantic scene completion. CVSformer (Dong et al. 2023) presents a multi-view feature synthesizer along with a cross-view Transformer that uses rotational kernels to generate multi-view voxel features and integrate their complementary information, greatly improving 3D semantic reasoning in occluded areas. In addition, VPNet (Wang et al. 2024b) introduces a confidence voxel generation mechanism combined with multi-frame knowledge distillation.

However, voxel-based approaches face an inherent trade-off between resolution and memory consumption, resulting in substantial computational resource demands in large-scale or high-precision scenarios. To improve efficiency while preserving structural details, point clouds, as a lightweight and flexible 3D representation, have gradually emerged as a new research direction in SSC (He et al.

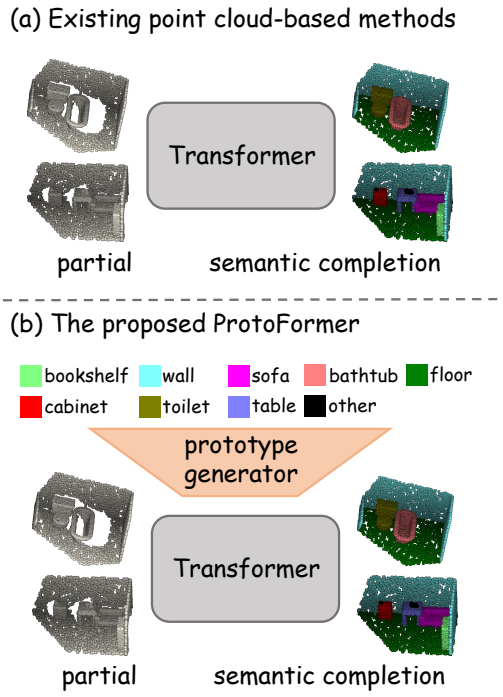

Figure 1: An illustration of our core idea. Subfigures (a) and (b) illustrate existing point cloud-based methods and ProtoFormer. ProtoFormer encodes semantic information into semantic prototypes to guide the Transformer.

*Corresponding author.

2025). PCSSC-Net (Zhang et al. 2021) is the first to develop a semantic scene completion method specifically designed for point cloud objects. By employing a hierarchical encoder and a semantics-guided decoder, it effectively enhances the completion and segmentation performance in indoor scenes. Building upon this, CasFusionNet (Xu et al. 2023) further introduces a modular cascaded architecture achieving fine-grained scene completion at the point cloud level. PointSSC (Yan et al. 2024) proposes a spatial-aware Transformer combined with a completion and segmentation cooperative module to effectively enhance the representation capability for semantic understanding of 3D point clouds. Although the aforementioned methods achieve promising performance and are more efficient than voxel-based approaches, they do not fully exploit and utilize semantic information.

To address this challenge, we propose ProtoFormer, which encodes semantic information into learnable semantic prototypes to guide the Transformer for semantic scene completion. Subfigure (a) illustrates existing point cloud-based methods, which typically involve a Transformer-based model for point cloud semantic scene completion. Subfigure (b) illustrates our ProtoFormer, which encodes semantic information into semantic prototypes to guide the Transformer in point cloud scene reconstruction and semantic awareness to fully exploit semantic information for point cloud semantic scene completion. First, we employ a Transformer-based encoder to encode the partial point cloud into a global representation. Next, a prototype generator is utilized to encode semantic information into learnable semantic prototypes. Finally, through stacked Transformer-based decoders, the incomplete point cloud is progressively reconstructed into a semantically completed result by integrating the semantic prototypes with the global representation. Extensive qualitative and quantitative experiments show that our model attains state-of-the-art performance with low complexity (improved by **81%** over CasFusionNet). The contributions of this paper are as follows.

- We propose ProtoFormer, which encodes semantic information into learnable prototypes to guide the Transformer for point cloud semantic scene completion.
- In ProtoFormer, we utilize a top-K attention mechanism to simultaneously perform point cloud coordinate prediction and per-point semantic label prediction, while reducing the model complexity.
- Extensive quantitative and qualitative experiments demonstrate that ProtoFormer outperforms state-of-the-art approaches at extremely low complexity compared to existing point cloud-based methods.

Related Work

Semantic Scene Completion

Voxel-based methods. SSCNet (Song et al. 2017) is the first to propose the task of simultaneously performing 3D voxel geometry completion and semantic labeling. CVSformer (Dong et al. 2023) introduces a multi-view feature synthesizer and a cross-view Transformer, employing rotational kernels to generate multi-view voxel features and fuse their complementary information, significantly enhancing 3D semantic reasoning capabilities in occluded regions. VoxFormer (Li et al. 2023b) is a Transformer-based two-stage sparse voxel querying method that effectively estimates complete 3D voxel semantic scenes from single or multiple 2D images. VPNet (Wang et al. 2024b) proposes a confidence voxel generation mechanism combined with multi-frame knowledge distillation. SidePaint (Huan et al. 2024) innovatively decomposes 3D scene completion into multiple side-view 2D contextual restoration sub-tasks and introduces distance-aware projection to generate dense voxel features, effectively reducing the complexity of 3D spatial reasoning and improving completion accuracy. CGFormer (Yu et al. 2024c) is a voxel Transformer that integrates context-aware query generation with 3D geometry awareness by lifting 2D image features into the 3D voxel space and fusing multiple 3D representations, effectively addressing depth ambiguity and feature aggregation issues.

Point cloud-based methods. PCSSC-Net (Zhang et al. 2021) is the first method specifically developed for semantic scene completion of point cloud objects. It utilizes a hierarchical encoder alongside a semantics-guided decoder to significantly improve completion and segmentation performance in indoor scenes. Building on this foundation, CasFusionNet (Xu et al. 2023) introduces a modular cascaded architecture that achieves fine-grained scene completion at the point cloud level. Meanwhile, PointSSC (Yan et al. 2024) presents a spatial-aware Transformer paired with a cooperative completion and segmentation module, effectively enhancing the representation capability for semantic understanding of 3D point clouds.

Single-Object Point Cloud Completion

Point cloud completion (Yuan et al. 2018; Tesema et al. 2024) aims to reconstruct a complete 3D object from a single partial observation, which may result from sensor limitations or environmental factors. Transformer (Vaswani et al. 2017) has been widely applied to point cloud completion (Yu et al. 2021; Zhou et al. 2022; Aiello, Valsesia, and Magli 2022; Chen et al. 2023; Li et al. 2023a; Zhu et al. 2023; Xiang et al. 2023; Yu et al. 2023; Zhang et al. 2023; Duan, Yu, and Chen 2024; Yu et al. 2024b; Xu et al. 2024; Yu et al. 2024a; Wang et al. 2024a; Rong et al. 2024; Zhong et al. 2025; Fang et al. 2025) due to its powerful global modeling capability. SnowflakeNet (Xiang et al. 2023) designs a snowflake point deconvolution to simulate the process of increasing point cloud resolution. AdaPoinTr (Yu et al. 2023) proposes an adaptive geometry-aware Transformer with denoising capabilities for point cloud completion. PointAttN (Wang et al. 2024a) constructs a fully attention-based network to achieve dense point cloud generation. SymmCompletion (Yan et al. 2025) designs a Transformer network based on symmetry priors for point cloud completion.

Although the aforementioned methods have achieved good performance in point cloud completion, they are not effective at handling complex semantic scenes. In the experimental section, we also compare several representative Transformer-based single-object point cloud completion methods from both quantitative and qualitative perspectives.

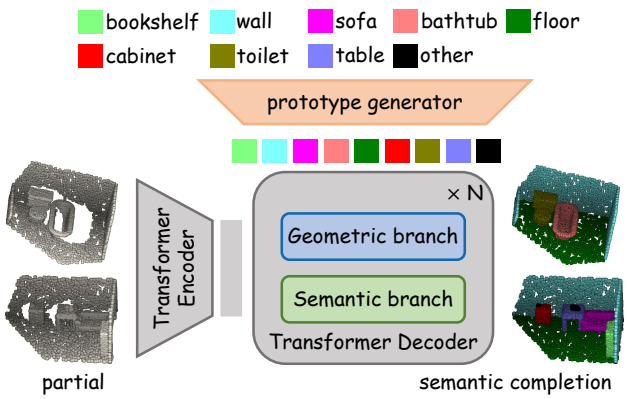

Figure 2: Overview of the proposed ProtoFormer. First, we encode the partial point cloud into a global representation using a Transformer-based encoder. Subsequently, the semantic information is encoded into sets of semantic prototypes. Finally, from both geometric and semantic perspectives, semantic prototypes are leveraged to guide multiple Transformer-based decoders in generating the final semantic completion results.

Method

Problem Statement

Given a partially observed scene point cloud set $P = \{p_i \mid i = 1, 2, \ldots, M\}$, where each point $p_i \in \mathbb{R}^3$. The goal of point cloud semantic scene completion is to predict a complete point cloud set while assigning a corresponding semantic label to each point. Specifically, we learn a mapping function

$$f : P \rightarrow \{\hat{P}, S\}, \tag{1}$$

where $\hat{P} = \{\hat{p}_i \mid i = 1, 2, \ldots, \hat{M}\}$ denotes the predicted scene point cloud. $S = \{s_i \mid i = 1, 2, \ldots, \hat{M}\}$ denotes each point's semantic label, where $s_i \in \{1, 2, \ldots, C\}$ and C is the predefined number of object categories.

Overview

Figure 2 illustrates the overall architecture of the proposed ProtoFormer. Initially, the partial point cloud is encoded into a global representation through a Transformer-based encoder. A prototype generator is utilized to encode semantic information into a set of distinct semantic prototypes, each corresponding to different object categories such as bookshelf, wall, sofa, bathtub, floor, cabinet, toilet, table, and others. Subsequently, multiple Transformer-based decoders, each comprising separate geometry and semantic branches, refine the representation and semantic prototypes across N decoding layers. This process ultimately produces the final semantic completion of the partial point cloud, effectively recovering both geometric structure and semantic labels.

Prototype-Guided Transformer

In this section, we will provide a detailed presentation to ProtoFormer. Generally, we denote the input point cloud as $P \in \mathbb{R}^{N\times 3}$. First, we pass it through an Transformer-based encoder, similar to the encoder used in SnowflakeNet (Xiang et al. 2023), to obtain a global feature vector $g \in \mathbb{R}^{1\times d}$.

Prototype Generator. This section describes how we encode semantic information into prototypes, treating them as priors to guide the network in semantic completion. Let the number of semantic categories be denoted as C. First, we generate C one-hot vectors, which together form an identity matrix $I \in \mathbb{R}^{C\times C}$. Then, we use MLP to encode the identity matrix I into coarse semantic prototypes, the formula as

$$T_c = \mathrm{MLP}(I), \tag{2}$$

where $T_c \in \mathbb{R}^{N_c\times C}$ is the coarse semantic prototypes and N_c denotes the dimension of prototypes.

Consistent with previous methods (Xu et al. 2023), we use farthest point sampling (Qi et al. 2017b) on a subset of the point cloud P to obtain a coarse point cloud $P_c \in \mathbb{R}^{N_c\times 3}$. Then, we concatenate the coarse point cloud P_c with the global vector g and pass them through an MLP to obtain coarse per-point semantic prediction results, the formula as

$$L_c = \mathrm{MLP}(P_c||\mathrm{tile}(g)), \tag{3}$$

where L_c is the predicted per-point semantic label of the coarse point cloud P_c.

Next, we introduce the working principle of the Transformer-based decoder. Generally, we assume the inputs are a low-resolution point cloud $P_{\mathrm{in}} \in \mathbb{R}^{N_{\mathrm{in}}\times 3}$, a low-resolution semantic prediction label $L_{\mathrm{in}} \in \mathbb{R}^{N_{\mathrm{in}}\times C}$, coarse semantic prototypes $T_{\mathrm{in}} \in \mathbb{R}^{N_{\mathrm{in}}\times C}$, and a global vector $g \in \mathbb{R}^{1\times d}$, while the outputs are a high-resolution point cloud $P_{\mathrm{out}} \in \mathbb{R}^{N_{\mathrm{out}}\times 3}$, a high-resolution semantic prediction label $L_{\mathrm{out}} \in \mathbb{R}^{N_{\mathrm{out}}\times C}$, and fine semantic prototypes $T_{\mathrm{out}} \in \mathbb{R}^{N_{\mathrm{out}}\times C}$. The formula as

$$P_{\mathrm{out}}, L_{\mathrm{out}}, T_{\mathrm{out}} = \mathrm{TD}(P_{\mathrm{in}}, L_{\mathrm{in}}, T_{\mathrm{in}}, g), \tag{4}$$

where $\mathrm{TD}(\cdot)$ is a Transformer-based decoder. Next, we introduce the two important branches in the Transformer-based decoder (TD), namely the geometric branch and the semantic branch, which are used to predict the high-resolution point cloud and the high-resolution semantic labels, respectively.

The previously generated semantic prototypes play an information enhancement role during the decoding phase, guiding the geometric and semantic branches to respectively produce high-resolution point clouds and high-resolution semantic labels.

Geometric Branch. First, the low-resolution point cloud $P_{\mathrm{in}} \in \mathbb{R}^{N_{\mathrm{in}}\times 3}$, combined with the coarse semantic prototypes $T_{\mathrm{in}} \in \mathbb{R}^{N_{\mathrm{in}}\times C}$ and global vectors $g \in \mathbb{R}^{1\times d}$, is used to generate the key (K) and value (V). The formula as

$$\begin{cases} Q = P_{\mathrm{in}}, \\ K = \mathrm{PointNet}(P_{\mathrm{in}}, g, T_{\mathrm{in}}), \\ V = \mathrm{PointNet}(P_{\mathrm{in}}, g, T_{\mathrm{in}}), \end{cases} \tag{5}$$

where PointNet is a lightweight version of (Qi et al. 2017a). Then, a high-resolution point cloud is generated through the top-K attention mechanism, the formula as

$$P_{\mathrm{out}} = \mathrm{tile}(P_{\mathrm{in}}) + \mathrm{DeConv}(\mathrm{Atte}(Q, K, V)), \tag{6}$$

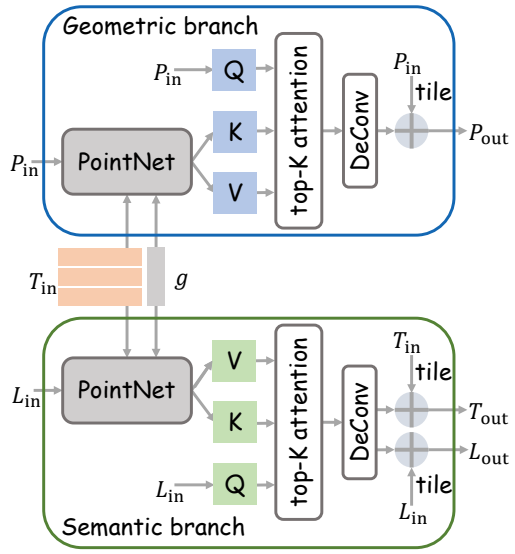

Figure 3: Overview of the proposed ProtoFormer decoder, which primarily comprises two branches, i.e., the geometric branch and the semantic branch. Here, T_{in} and T_{out} denote the coarse and fine semantic prototypes.

where $\text{Atte}(\cdot)$ denotes the top-K attention mechanism, $\text{DeConv}(\cdot)$ is the deconvolution operation, and $\text{tile}(\cdot)$ is the copy operation. $P_{\text{out}} \in \mathbb{R}^{\mu N_{\text{in}} \times 3}$ is the high-resolution point cloud and μ is the upsampling rate.

Semantic Branch. The principle of the semantic branch is similar to the aforementioned process. First, the low-resolution semantic labels $L_{\text{in}} \in \mathbb{R}^{N_{\text{in}} \times C}$, combined with the coarse semantic prototypes T_{in} and global vectors g, is used to generate the key (K) and value (V). The formula as

$$\begin{cases} Q = L_{\text{in}}, \\ K = \text{PointNet}(L_{\text{in}}, g, T_{\text{in}}), \\ V = \text{PointNet}(L_{\text{in}}, g, T_{\text{in}}). \end{cases} \tag{7}$$

Then, a high-resolution point cloud is generated and fine semantic prototypes through the top-K attention mechanism, the formula as

$$\begin{cases} L_{\text{out}} = \text{tile}(L_{\text{in}}) + \text{DeConv}(\text{Atte}(Q, K, V)), \\ T_{\text{out}} = \text{tile}(T_{\text{in}}) + \text{DeConv}(\text{Atte}(Q, K, V)), \end{cases} \tag{8}$$

where $L_{\text{out}} \in \mathbb{R}^{\mu N_{\text{in}} \times C}$ and $T_{\text{out}} \in \mathbb{R}^{\mu N_{\text{in}} \times C}$ are the high-resolution point cloud and fine semantic prototypes, respectively. μ is the upsampling rate. **From this part, we can observe that our semantic prototypes are learnable, and as the resolution of the point cloud increases, each semantic prototype contains more information.**

In the following, we present the operational principle of the top-K attention mechanism.

Top-K Attention Mechanism. Generally, the input query, key, and value are denoted as $Q = \{q_i\}_{i=1}^{N}$, $K = \{k_i\}_{i=1}^{N}$, and $V = \{v_i\}_{i=1}^{N}$, respectively. Cosine similarity is computed between each query and key vector pair to generate a similarity matrix. From this matrix, the top K most relevant key vectors are selected for each query vector, resulting in a sparser similarity matrix that serves as the adjacency matrix. The edges connecting the i-th query vector to its top K most relevant value vectors are then defined as

$$a_{ij} = \frac{q_i \cdot {k_j}^T}{\|q_i\|_2 \cdot \|k_j\|_2}, j = 1, 2, \ldots, \text{K}. \tag{9}$$

Finally, we use this sparse adjacency matrix $A = [a_{ij}]$ to perform the aggregation of the values.

$$h_i = \sum_{j=1}^{\text{K}} (a_{ij} \cdot v_j), i = 1, 2, \ldots, N, \tag{10}$$

$H = \{h_i\}_{i=1}^{N}$ is the output of the top-K attention mechanism. In summary, ProtoFormer generates the final semantic completion result by performing multiple iterations of formula (4).

Overall, ProtoFormer utilizes learnable semantic prototypes to guide both the geometric and semantic branches, enabling the generation of more refined semantic completion results.

Loss

Similar to previous methods (Xu et al. 2023), we train ProtoFormer using a two-part loss function, i.e., a Chamfer distance (CD) (Fan, Su, and Guibas 2017) loss to constrain the generated point cloud coordinates, and a semantic loss to supervise the per-point semantic labels of the completion results. The formula of CD is as

$$\begin{aligned} \mathcal{L}_{\text{CD}}(P, \hat{P}) = &\frac{1}{|P|} \sum_{p \in P} \min_{\hat{p} \in \hat{P}} \|p - \hat{p}\|_2 \\ &+ \frac{1}{|\hat{P}|} \sum_{\hat{p} \in \hat{P}} \min_{p \in P} \|\hat{p} - p\|_2, \end{aligned} \tag{11}$$

where P and $\hat{P}$ are the completion result and ground truth, respectively. Since point clouds are unstructured, for each predicted point, we find its nearest point in the ground truth point cloud using FPS (Qi et al. 2017b) and assign the corresponding semantic label to compute the semantic loss. The specific formulation is as

$$\mathcal{L}_{\text{seg}} = \sum_{l \in L} -w_l \log(l), \tag{12}$$

where $l \in L$ is predict labels gathered from ground-truth labels $\hat{l} \in \hat{L}$ and w_l is the l-th label's corresponding weight used to mitigate class imbalance across different categories.

Finally, the total loss as

$$\mathcal{L} = \sum_{i=1}^{N} (\mathcal{L}_{\text{CD}}(P_i, \hat{P}_i) + \alpha \mathcal{L}_{\text{seg}}(L_i, \hat{L}_i)), \tag{13}$$

where N is the number of Transformer-based decoder.

Methods	CD ↓	F-Score ↑	mAcc ↑	mIoU ↑	Params (M) ↓	Model Size (MB) ↓
Single-object point cloud completion methods						
SnowflakeNet (TPAMI 2023)	13.120	0.335	–	–	19.30	221.14
AdaPoinTr (TPAMI 2023)	16.071	0.287	–	–	32.46	371.36
PointAttN (AAAI 2024)	14.603	0.333	–	–	29.79	341.23
SymmCompletion (AAAI 2025)	15.115	0.320	–	–	13.28	149.87
Point cloud semantic scene completion methods						
PointSSC (ICRA 2024)	18.827	0.223	79.23	71.92	20.76	237.31
CasFusionNet (AAAI 2023)	9.325	0.545	94.93	**91.65**	19.85	227.71
ProtoFormer (Ours)	**8.917**	**0.559**	**95.48**	90.71	**3.75** (**−81.1%**)	**43.25** (**− 81.0%**)

Table 1: Quantitative experimental comparison of different methods on SSC-PC.

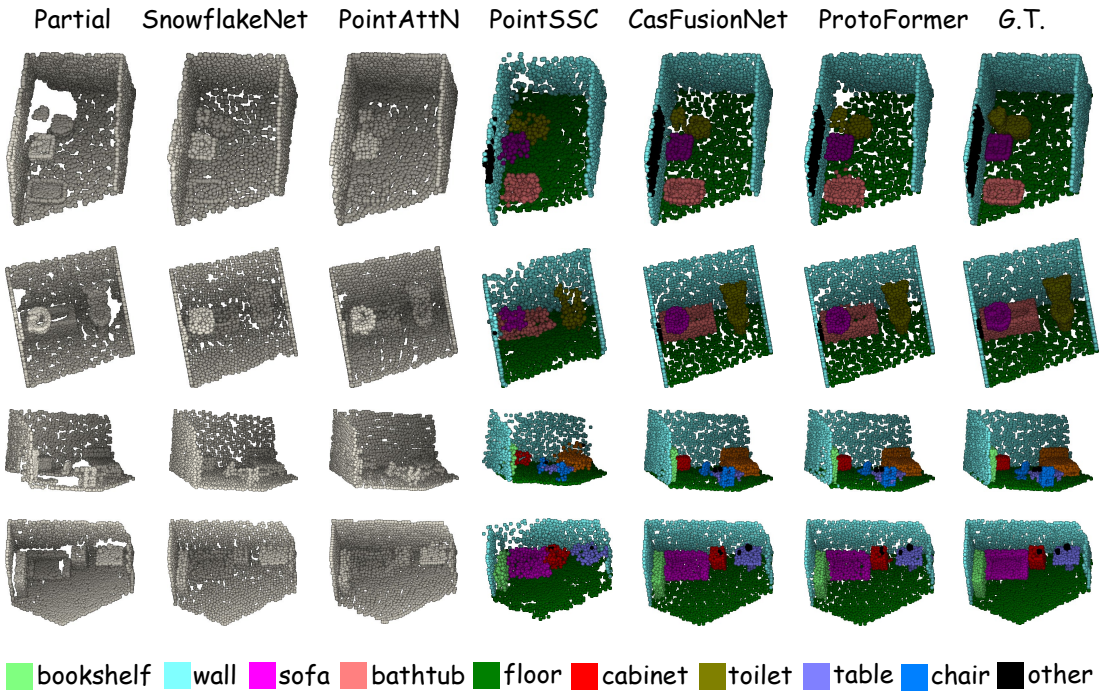

Figure 4: Qualitative experimental comparison of different methods on SSC-PC.

Experiments and Analyses

Datasets

Following previous work (Xu et al. 2023), we conducted experiments on two datasets, i.e., SSC-PC and NYUCAD-PC.

SSC-PC. SSC-PC contains four scenes, i.e., bathroom, bedroom, living room, and office, with a total of 1,941 samples covering 16 object categories. Following the same data split as the previous method (Xu et al. 2023), we use 1,543 samples as the training set and the remaining 398 samples as the test set. The input point cloud resolution for each sample is 4,096 points, and the output point cloud resolution is also 4,096 points, with each point containing a semantic category label.

NYUCAD-PC. NYUCAD-PC is a more complex point cloud semantic scene completion dataset containing 11 object categories and covering scenes such as bathroom, bedroom, living room, office, bookstore, kitchen, study room, classroom, computer lab, foyer, home office, office kitchen, playroom, reception room, and dining room. Using the same data split as the previous method (Xu et al. 2023), 795 samples are designated for training, while the remaining 654 samples are used for testing. Each sample's input point cloud has a resolution of 4,096 points, and the output point cloud has a resolution of 8,192 points, with each point annotated with a category label.

Experimental Settings

Metrics. Similar to previous methods (Xu et al. 2023; Yan et al. 2024), we evaluate the quality of the semantic comple-

Methods	CD ↓	F-Score ↑	mAcc ↑	mIoU ↑	Params (M) ↓	Model Size (MB) ↓
Single-object point cloud completion methods						
SnowflakeNet (TPAMI 2023)	12.139	0.683	–	–	19.30	221.23
AdaPoinTr (TPAMI 2023)	13.859	0.568	–	–	32.47	371.50
PointAttN (AAAI 2024)	11.874	0.647	–	–	31.43	360.01
SymmCompletion (AAAI 2025)	13.964	0.523	–	–	13.28	149.88
Point cloud semantic scene completion methods						
PointSSC (ICRA 2024)	15.834	0.534	46.64	36.76	20.76	237.33
CasFusionNet (AAAI 2023)	**10.173**	**0.765**	59.57	**49.33**	25.80	296.01
ProtoFormer (Ours)	10.673	0.762	**60.68**	49.19	**4.51** (− **82.5%**)	**52.06** (− **82.4%**)

Table 2: Quantitative experimental comparison of different methods on NYUCAD-PC.

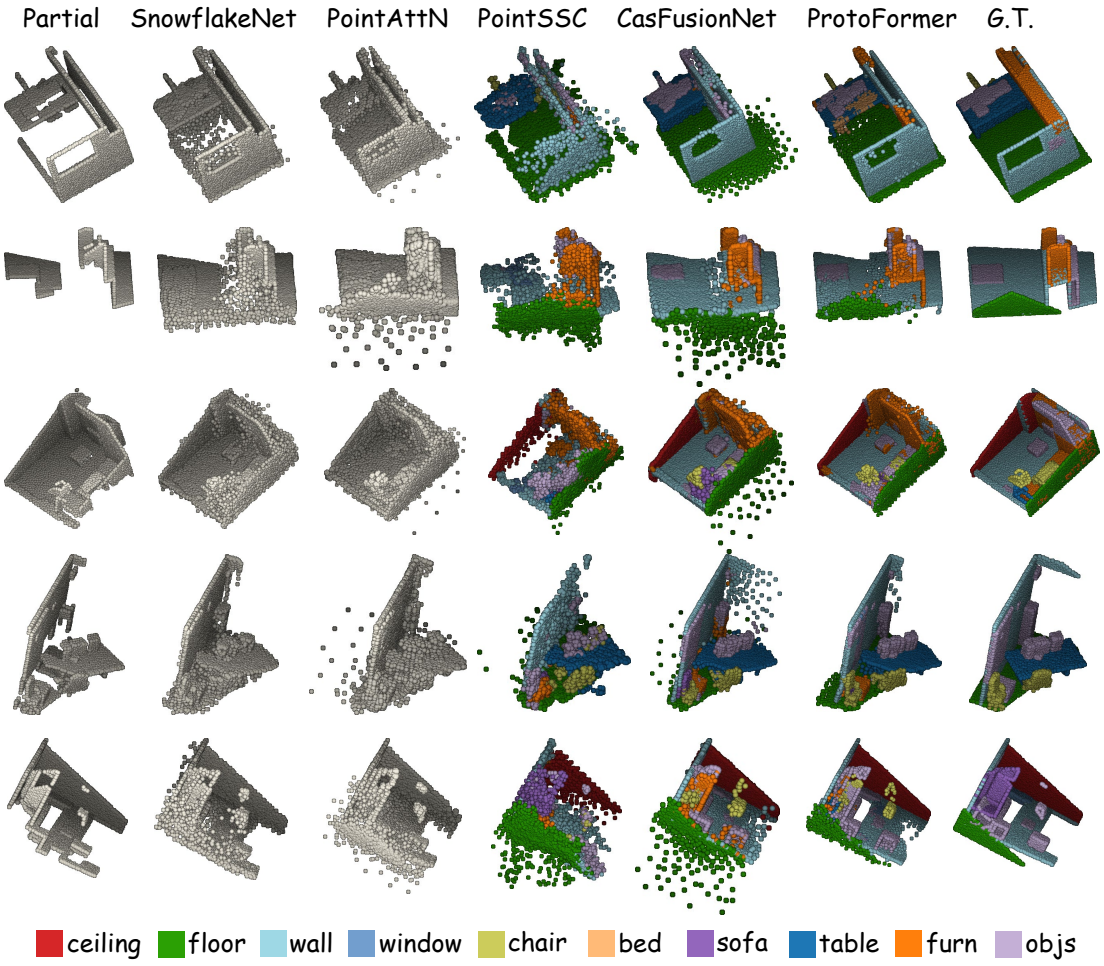

Figure 5: Qualitative experimental comparison of different methods on NYUCAD-PC.

tion results using commonly adopted metrics in point cloud completion tasks, such as CD and F-Score, as well as metrics frequently used in semantic segmentation, including mAcc and mIoU.

Implementation Details. For both the SSC-PC and NYUCAD-PC datasets, we train the model using the Adam optimizer. Our experiments were conducted on an NVIDIA H100. Specifically, the batch size is set to 8, the number of epochs to 400, and the initial learning rate to 0.0005. The learning rate is decayed by a factor of 0.9 every 100 epochs. For SSC-PC, ProtoFormer employs two Transformer decoders with upsampling rates of 1 and 2, respectively, along

Methods	CD ↓	F-Score ↑	mAcc ↑	mIoU ↑	Params (M) ↓	Model Size (MB) ↓	Memory (MB) ↓	Inference Time (s) ↓
SSC-PC								
CasFusionNet (AAAI 2023)	9.325	0.545	94.93	**91.65**	19.85	227.71	3906	28
ProtoFormer (Ours)	**8.917**	**0.559**	**95.48**	90.71	**3.75** (− **81.1%**)	**43.25** (− **81.0%**)	**2002** (− **48.7%**)	**10** (− **64.3%**)
NYUCAD-PC								
CasFusionNet (AAAI 2023)	**10.173**	**0.765**	59.57	**49.33**	25.80	296.01	7382	87
ProtoFormer (Ours)	10.673	0.762	**60.68**	49.19	**4.51** (− **82.5%**)	**52.06** (− **82.4%**)	**3490** (− **52.7%**)	**35** (− **59.8%**)

Table 3: Complexity comparison with the state-of-the-art method CasFusionNet on SSC-PC and NYUCAD-PC.

with a balancing weight w_l is set to $\{1.50, 0.96, 1.03, 1.10, 1.67, 1.10, 1.14, 1.74, 0.69, 1.06, 2.09, 1.10, 1.06, 1.57, 1.68, 0.69\}$ for the semantic loss. For NYUCAD-PC, ProtoFormer utilizes three Transformer decoders, each with an upsampling rate of 2 for the semantic loss.

Evaluation on SSC-PC

Quantitative Experiments. We compare several advanced single-object point cloud completion methods such as SnowflakeNet (Xiang et al. 2023), AdaPoinTr (Yu et al. 2023), PointAttN (Wang et al. 2024a), and SymmCompletion (Yan et al. 2025), as well as open-source point cloud semantic scene completion methods like CasFusionNet (Xu et al. 2023) and PointSSC (Yan et al. 2024). Table 1 presents a comparison of the quantitative experimental results. It can be observed that our method attains performance comparable to that of CasFusionNet, while simultaneously achieving an **81.1%** reduction in the parameter count.

Qualitative Experiments. We compare several state-of-the-art single-object point cloud completion methods, including SnowflakeNet (Xiang et al. 2023) and PointAttN (Wang et al. 2024a), alongside open-source semantic scene completion approaches such as CasFusionNet (Xu et al. 2023) and PointSSC (Yan et al. 2024). Figure 4 presents a comparison of qualitative experimental results, showing that single-object point cloud completion methods struggle to perceive different semantic objects due to the lack of guidance from semantic information. Compared to other point cloud semantic scene completion methods, our approach achieves performance comparable to state-of-the-art techniques in recovering missing geometric shapes and perceiving different semantic objects.

Evaluation on NYUCAD-PC

Quantitative Experiments. We conduct a comparison of several advanced single-object point cloud completion methods, including SnowflakeNet (Xiang et al. 2023), AdaPoinTr (Yu et al. 2023), PointAttN (Wang et al. 2024a), and SymmCompletion (Yan et al. 2025), alongside open-source point cloud semantic scene completion methods such as CasFusionNet (Xu et al. 2023) and PointSSC (Yan et al. 2024). Table 2 presents a comparison of the quantitative experimental results. Our method achieves comparable performance to CasFusionNet while simultaneously reducing the number of parameters by **82.5%**.

Qualitative Experiments. Our qualitative comparison includes several single-object point cloud completion methods, such as SnowflakeNet (Xiang et al. 2023) and PointAttN (Wang et al. 2024a), as well as open-source point cloud semantic scene completion methods like CasFusionNet (Xu et al. 2023) and PointSSC (Yan et al. 2024). Given the heightened complexity of the NYUCAD-PC dataset, certain comparative methods, including CasFusionNet and PointAttN, tend to produce a considerable number of outliers. In contrast, our method demonstrates superior performance in both reconstructing missing regions and semantic label prediction.

Methods	CD ↓	F-Score ↑	mAcc ↑	mIoU ↑
SSC-PC				
ProtoFormer w/o prototypes	9.136	0.556	93.41	87.15
ProtoFormer (Ours)	**8.917**	**0.559**	**95.48**	**90.71**
NYUCAD-PC				
ProtoFormer w/o prototypes	11.133	0.710	50.08	38.92
ProtoFormer (Ours)	**10.673**	**0.762**	**60.68**	**49.19**

Table 4: Quantitative experiments of ablation study.

Complexity Analyses and Ablation Study

Complexity Analyses. Tables 1 and 2 present a comparison of the computational complexity between our method and existing approaches. It can be observed that our method achieves competitive semantic completion performance with significantly lower complexity. In addition, Table 3 presents a detailed complexity comparison between ProtoFormer and the state-of-the-art method CasFusionNet. The results demonstrate that our approach achieves significant improvements across multiple efficiency metrics.

Ablation Study. We conducted an ablation study on our core contribution, the semantic prototype. The quantitative results, presented in Table 4, demonstrate that the semantic prototype is crucial to the effectiveness of our method.

Conclusion

This paper proposes ProtoFormer, which encodes semantic information into learnable prototypes, enabling the Transformer to better predict semantic completion results. Specifically, within the Transformer decoder, we employ a top-K attention mechanism to achieve efficient information aggregation. Extensive qualitative and quantitative experiments demonstrate that ProtoFormer achieves state-of-the-art performance in semantic completion while maintaining remarkably low computational complexity.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U21A20473, 62376142, 62536006, 62032022) and the Key Technologies Program of Taihang Laboratory in Shanxi Province, China (THYF-JSZX-24010600).

References

Aiello, E.; Valsesia, D.; and Magli, E. 2022. Cross-modal learning for image-guided point cloud shape completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 37349–37362. New Orleans, USA.

Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; and Li, H. 2024. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12): 10164–10183.

Chen, Z.; Long, F.; Qiu, Z.; Yao, T.; Zhou, W.; Luo, J.; and Mei, T. 2023. AnchorFormer: Point cloud completion from discriminative nodes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13581–13590. Vancouver, Canada.

Cheng, Y.; Su, J.; Jiang, M.; and Liu, Y. 2022. A novel radar point cloud generation method for robot environment perception. *IEEE Transactions on Robotics (TRO)*, 38(6): 3754–3773.

Dong, H.; Ma, E.; Wang, L.; Wang, M.; Xie, W.; Guo, Q.; Li, P.; Liang, L.; Yang, K.; and Lin, D. 2023. CVSformer: Cross-view synthesis Transformer for semantic scene completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8840–8849. Paris, France.

Du, Z.; Liang, J.; Liang, J.; Yao, K.; and Cao, F. 2024. Graph regulation network for point cloud segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12): 7940–7955.

Duan, F.; Yu, J.; and Chen, L. 2024. T-CorresNet: Template guided 3D point cloud completion with correspondence pooling query generation strategy. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 99–106. Milan, Italy.

Fan, H.; Su, H.; and Guibas, L. J. 2017. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 605–613. Honolulu, USA.

Fang, C.; Liang, J.; Liang, J.; Wang, H.; Yao, K.; and Cao, F. 2025. Multi-modal point cloud completion with interleaved attention enhanced Transformer. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 963–971. Montreal, Canada.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3354–3361. Providence, USA.

He, W.; Chen, X.; Chen, W.; Wang, H.; Liu, Y.; and Li, R. 2025. RWKV-PCSSC: Exploring RWKV model for point cloud semantic scene completion. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 161–170. Dublin, Ireland.

Huan, L.; Dong, M.; Yue, L.; Shen, S.; and Zheng, X. 2024. Easing 3D pattern reasoning with side-view features for semantic scene completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 440–455. Milan, Italy.

Li, A.; Zimmer-Dauphinee, J. R.; Kalyanam, R.; Lindsay, I.; VanValkenburgh, P.; Wernke, S.; and Aliaga, D. 2025. Self-supervised large scale point cloud completion for archaeological site restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11759–11768. Nashville, USA.

Li, S.; Gao, P.; Tan, X.; and Wei, M. 2023a. ProxyFormer: Proxy alignment assisted point cloud completion with missing part sensitive Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9466–9475. Vancouver, Canada.

Li, Y.; Yu, Z.; Choy, C.; Xiao, C.; Alvarez, J. M.; Fidler, S.; Feng, C.; and Anandkumar, A. 2023b. VoxFormer: Sparse voxel Transformer for camera-based 3D semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9087–9098. Vancouver, Canada.

Liang, J.; Du, Z.; Liang, J.; Yao, K.; and Cao, F. 2023. Long and short-range dependency graph structure learning framework on point cloud. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(12): 14975–14989.

Liang, Y.; Chen, B.; and Song, S. 2021. SSCNav: Confidence-aware semantic scene completion for visual semantic navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 13194–13200. Xi'an, China.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 77–85. Honolulu, USA.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 5105–5114. Long Beach, USA.

Rong, Y.; Zhou, H.; Yuan, L.; Mei, C.; Wang, J.; and Lu, T. 2024. CRA-PCN: Point cloud completion with intra- and inter-level cross-resolution Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 4676–4685. Vancouver, Canada.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1746–1754. Honolulu, USA.

Tesema, K. W.; Hill, L.; Jones, M. W.; Ahmad, M. I.; and Tam, G. K. 2024. Point cloud completion: A survey. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 30(10): 6880–6899.

Varley, J.; DeChant, C.; Richardson, A.; Ruales, J.; and Allen, P. 2017. Shape completion enabled robotic grasping. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2442–2447. Vancouver, Canada.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 6000–6010. Long Beach, USA.

Wang, J.; Cui, Y.; Guo, D.; Li, J.; Liu, Q.; and Shen, C. 2024a. PointAttN: You only need attention for point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 5472–5480. Vancouver, Canada.

Wang, L.; Lin, D.; Yang, K.; Liu, R.; Guo, Q.; Xie, W.; Wang, M.; Liang, L.; Wang, Y.; and Li, P. 2024b. Voxel proposal network via multi-frame knowledge distillation for semantic scene completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 101096–101115. Vancouver, Canada.

Xiang, P.; Wen, X.; Liu, Y.-S.; Cao, Y.-P.; Wan, P.; Zheng, W.; and Han, Z. 2023. Snowflake point deconvolution for point cloud completion and generation with skip-Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(5): 6320–6338.

Xu, H.; Long, C.; Zhang, W.; Liu, Y.; Cao, Z.; Dong, Z.; and Yang, B. 2024. Explicitly guided information interaction network for cross-modal point cloud completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 414–432. Milan, Italy.

Xu, J.; Li, X.; Tang, Y.; Yu, Q.; Hao, Y.; Hu, L.; and Chen, M. 2023. CasFusionNet: A cascaded network for point cloud semantic scene completion by dense feature fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, 3018–3026. Washington, USA.

Yan, H.; Li, Z.; Luo, K.; Lu, L.; and Tan, P. 2025. SymmCompletion: High-fidelity and high-consistency point cloud completion with symmetry guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 9094–9102. Philadelphia, USA.

Yan, Y.; Liu, B.; Ai, J.; Li, Q.; Wan, R.; and Pu, J. 2024. PointSSC: A cooperative vehicle-infrastructure point cloud benchmark for semantic scene completion. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 17027–17034. Yokohama, Japan.

Yu, J.; Huang, B.; Zhang, Y.; Li, H.; Tang, X.; and Gao, S. 2024a. GeoFormer: Learning point cloud completion with tri-plane integrated Transformer. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 8952–8961. Melbourne, Australia.

Yu, X.; Rao, Y.; Wang, Z.; Liu, Z.; Lu, J.; and Zhou, J. 2021. PoinTr: Diverse point cloud completion with geometry-aware Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 12478–12487. Montreal, Canada.

Yu, X.; Rao, Y.; Wang, Z.; Lu, J.; and Zhou, J. 2023. AdaPoinTr: Diverse point cloud completion with adaptive geometry-aware Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(12): 14114–14130.

Yu, X.; Wang, Y.; Zhou, J.; and Lu, J. 2024b. ProtoComp: Diverse point cloud completion with controllable prototype. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 270–286. Milan, Italy.

Yu, Z.; Zhang, R.; Ying, J.; Yu, J.; Hu, X.; Luo, L.; Cao, S.-Y.; and Shen, H.-L. 2024c. Context and Geometry Aware Voxel Transformer for semantic scene completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 1531–1555. Vancouver, Canada.

Yuan, W.; Khot, T.; Held, D.; Mertz, C.; and Hebert, M. 2018. PCN: Point completion network. In *Proceedings of the International Conference on 3D Vision (3DV)*, 728–737. Verona, Italy.

Zhang, S.; Li, S.; Hao, A.; and Qin, H. 2021. Point cloud semantic scene completion from RGB-D images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 3385–3393. Virtual.

Zhang, W.; Zhou, H.; Dong, Z.; Liu, J.; Yan, Q.; and Xiao, C. 2023. Point cloud completion via skeleton-detail Transformer. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 29(10): 4229–4242.

Zhong, Y.; Quan, W.; Yan, D.-M.; Jiang, J.; and Wei, Y. 2025. PointCFormer: A relation-based progressive feature extraction network for point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 10689–10697. Philadelphia, USA.

Zhou, H.; Cao, Y.; Chu, W.; Zhu, J.; Lu, T.; Tai, Y.; and Wang, C. 2022. SeedFormer: Patch seeds based point cloud completion with upsample Transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 416–432. Tel Aviv, Israel.

Zhu, Z.; Chen, H.; He, X.; Wang, W.; Qin, J.; and Wei, M. 2023. SVDFormer: Complementing point cloud via self-view augmentation and self-structure dual-generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14508–14518. Paris, France.