Semantic Guided Part Relation-aware Network for Point Cloud Completion

Zhensheng Zhou[1], Jianqing Liang[1*], Jiye Liang[1], Zijin Du[2], Chenghao Fang[1]

[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 Engineering Science, University of Chinese Academy of Sciences

zhouzhensheng@sxu.edu.cn, liangjq@sxu.edu.cn, ljy@sxu.edu.cn, hetempest0302@163.com, chenghaofang1014@163.com

Abstract

The primary goal of 3D point cloud completion is to reconstruct complete and high-resolution point clouds from incomplete and low-resolution inputs. Although recent approaches have achieved satisfactory completion performance by incorporating additional images, there remains substantial room for improvement in fully harnessing the rich geometric relational information inherent in the parts. To address this challenge, we propose a novel Semantic Guided Part Relation-aware Network (SGPRNet) for Point Cloud Completion. Its core innovation lies in establishing part semantic relations to guide the reconstruction of structurally consistent local geometries. Specifically, we utilize Multi-modal Large Language Models (MLLMs) to automatically generate the specific text of 3D shapes, which provides detailed descriptions of geometric part relations. Building upon this, we design an Orthogonal Semantic Part Transfer (OSPT) module that learns transferable semantic relations between geometric parts. Subsequently, we develop a Semantic Geometric Relation-aware Transformer (SGRFormer) to progressively refine these semantic features, enhancing point cloud representations and guiding the generation of fine local structures. In addition, we establish a point-text pairs corpus, OmniObject3D-212/34 and Text-ViPC datasets based on existing OmniObject3D and ShapeNet-ViPC datasets, incorporating the specific text. Extensive experimental results demonstrate that our method outperforms existing state-of-the-art completion methods.

Introduction

The recent rapid development of 3D sensors has driven extensive research in 3D computer vision. Point clouds are becoming increasingly important in real-world applications, such as robot navigation (Varley et al. 2017; Xiong et al. 2023), autonomous driving (Wang et al. 2023), and scene understanding (Hou, Dai, and Nießner 2019; Cao et al. 2024). However, point cloud data obtained from 3D sensors are usually incomplete, due to unavoidable self-occlusion, light reflection, and limited sensor resolution. Therefore, recovering complete point cloud data from partial and sparse raw data becomes an indispensable task with increasing importance for point cloud analysis (Liang et al. 2023; Du et al. 2024b; Zheng et al. 2021).

In recent years, the development of 3D point cloud completion has been driven by advances in deep learning technology. Existing single-modal point cloud completion methods primarily fall into two paradigms: voxelization-based and point-based methods. Both paradigms employ an encoder to extract a latent representation from partial input, however, they diverge primarily in their approach to decoding: voxelization-based approaches (Dai, Ruizhongtai Qi, and Nießner 2017; Stutz and Geiger 2018) utilize 3D convolutions to reconstruct complete shapes from latent representations, but at the cost of exponential computational complexity; point-based methods (Cui et al. 2023; Li et al. 2023; Lyu et al. 2021; Zhang, Yan, and Xiao 2020; Zhu et al. 2023a; Yuan et al. 2018; Xiang et al. 2022; Yu et al. 2021; Zhou et al. 2022) usually adopt an encoder-decoder pipeline to reconstruct fine-grained coarse point clouds from latent representations.

Unfortunately, although the above methods have achieved satisfactory results, it is difficult to reconstruct complete detailed geometric structures from sparse and disordered point clouds relying solely on point cloud input, which has become an ill-posed problem (Han, Laga, and Bennamoun 2019). To this end, the seminal work ViPC (Zhang et al. 2021) introduces additional images to supplement prior information for missing regions. Inspired by this, subsequent methods such as XMFNet (Aiello, Valsesia, and Magli 2022) and CSDN (Zhu et al. 2023b) leverage stacked cross-attention layers to extract global features from images, as illustrated in Figure 1. Nevertheless, the inherent limitations of the image modality make it difficult to align with and effectively capture the intrinsic geometric relations between structural parts.

Although images can provide global features of 3D shapes, they only present a single perspective, and therefore naturally lack sufficient depth information and multi-perspective spatial semantics. This makes it difficult for images to adequately describe the complex geometric relations between different parts of complex 3D shapes. This limitation is particularly evident in cases of severe incompleteness, which causes geometric ambiguity at part junctions. In contrast, text excels at defining these geometric structural relations and can encode the rich multi-perspective semantic information of an object's parts. Humans can interpret rich text

*Corresponding author.

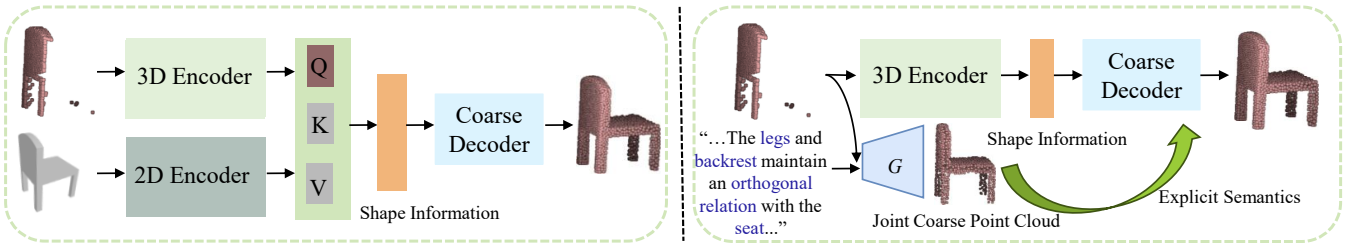

Figure 1: Existing shape completion methods (Left) implicitly decode part semantics from shape information. Our pipeline (Right) explicitly decodes them, which guarantees the generation of a joint coarse point cloud.

description (e.g.,"this is a symmetrical high-back chair with four legs perpendicular to the seat") as structured spatial prior information, enabling us to construct complete and reasonable 3D shapes in our minds even in the absence of specific geometric details. This capability inspires us to consider: can the structured spatial knowledge embedded in rich text descriptions be effectively leveraged to guide and enhance the point cloud completion process, particularly when involving severe incompleteness or unseen categories?

To overcome the aforementioned challenges, considering that manually constructing the specific text for each point cloud input is both costly and highly subjective, we introduce the MLLMs to automatically generate detailed 3D geometric and appearance descriptions as part semantic, thereby enhancing the relations between geometric parts. We transform the complex point cloud completion problem into a semantic recovery and geometric recovery. In the semantic recovery stage, the OSPT module employs a part relation dictionary to model geometric relations, compensating for the contextual semantic loss between incomplete geometric parts. Furthermore, in the geometric recovery stage, due to the inherent ambiguity and incompleteness of the input point cloud, relying solely on the incomplete input would lead the decoder to generate a coarse point cloud, which is rough and ambiguous. Therefore, in the geometric recovery stage, we introduce part semantic prompts. We treat these prompts as coarse part semantic features transferred via the SGRFormer module, and these features enhance the high consistency of geometric embedding representations. This ensures the generation of high fidelity coarse point clouds, which outline the geometric shape of the target object and guide the model to generate more fine and complete point clouds in a coarse-to-fine manner. The main contributions can be summarized in the following three folds:

- We propose a novel SGPRNet that leverages text generated by MLLMs as part semantic prompts to explicitly guide the network in modeling geometric part relations, enabling a global understanding of 3D shapes.
- An OSPT module is designed that leverages a learnable part relation dictionary to generate coarse part semantic features, compensating for the absence of semantic relations between incomplete geometric parts.
- Preserving the raw geometric structures of the input part point clouds, our SGRFormer module generates joint coarse point clouds, which in turn enhances the model's capability to construct fine and complete point clouds.

Related Work

Single-Modal Point Cloud Completion

TopNet (Tchapmi et al. 2019) implicitly encodes point clouds into a rooted tree topology and progressively reconstructs shapes through tree-structured decoding with hierarchical upsampling. PCN (Yuan et al. 2018) pioneered the coarse-to-fine point cloud completion paradigm, which decodes global features solely via multi-layer perceptrons (MLPs) to generate complete point clouds. While widely adopted for its simplicity (Huang et al. 2020; Xie et al. 2020; Wang, Ang Jr, and Lee 2020; Zhang, Yan, and Xiao 2020; Liu et al. 2020), vanilla MLPs struggle to capture complex geometric structures. To address this, SnowflakeNet (Xiang et al. 2022) introduces a snowflake point deconvolution decoder: at the coarse stage, it generates sparse point clouds through per-point MLPs operation on global features; during refinement, a cascaded Skip-Transformer fuses current-step features with historical ones, leveraging prior context to guide displacement prediction for optimized point generation. In contrast, Pointr (Yu et al. 2021) abandons the two-stage scheme by framing completion as a sequence-to-sequence translation task that directly outputs missing regions. Similar concepts later emerged in (Duan, Yu, and Chen 2024; Ying et al. 2023). Critically, prevailing two-stage methods overlook the pivotal impact of coarse seed point clouds quality on detailed reconstruction, thereby limiting fine-grained geometric recovery.

Multi-Modal Point Cloud Completion

The unstructured and sparse nature of point clouds limits information density, making it challenging to infer 3D spatial structures of missing regions—thus framing point cloud completion as the ill-posed completion task. In contrast, images offer more intuitive visual cues and richer global semantics. ViPC (Zhang et al. 2021) pioneered the use of global image priors to assist missing part reconstruction, yet its avoidance of explicit image-to-point cloud conversion overlooks inherent geometric characteristics, failing to deliver critical local structures. While existing Transformer based methods (Aiello, Valsesia, and Magli 2022; Fang et al. 2025) enable feature interaction in local latent spaces via cross-attention, it lacks image-point cloud feature alignment. CSDN (Zhu et al. 2023b) consolidates multimodal

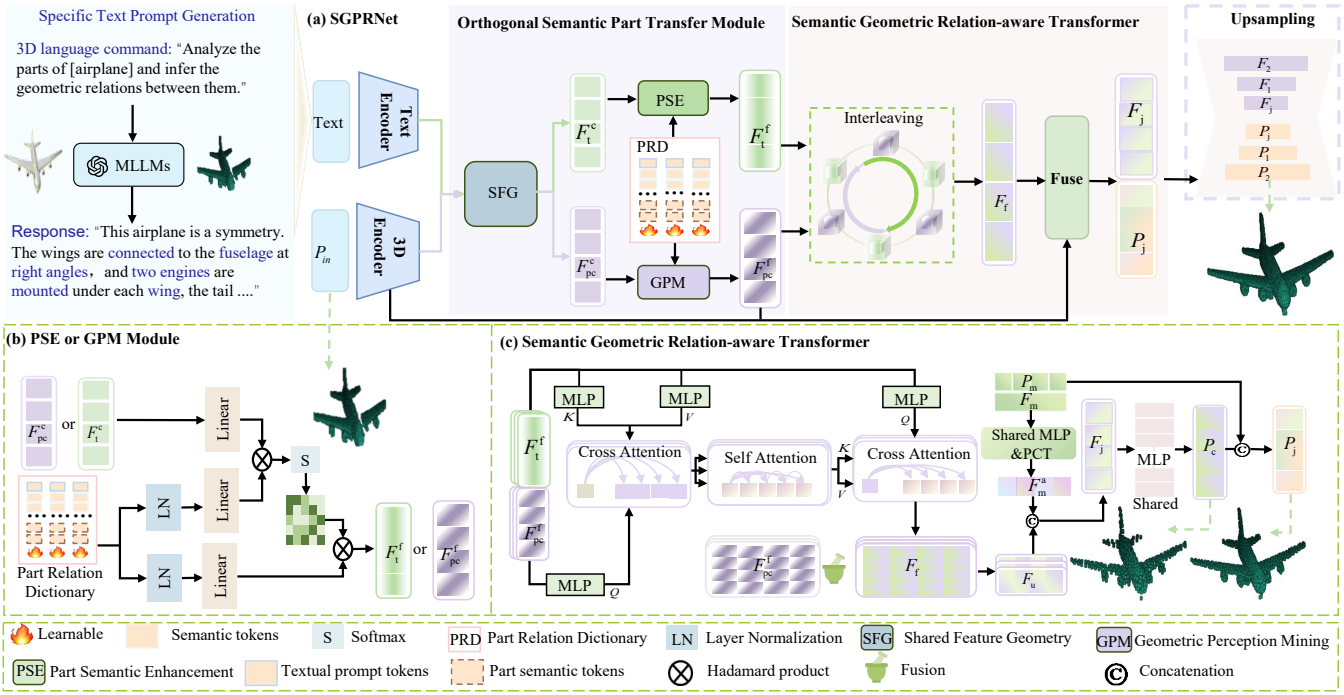

Figure 2: An overview of the proposed SGPRNet. In the orthogonal semantic part transfer stage, we learn transferable semantic relations between geometric parts based on semantic prompts. These relations are subsequently integrated with the partial point clouds to reconstruct the joint coarse point cloud in the semantic geometric relation-aware transformer stage.

features into unidirectional hybrid representations, but its performance degrades under severe occlusion due to single-view depth ambiguity and contextual insufficiency. Text descriptions provide an alternative pathway: their rich shape semantics enhance network comprehension of partial structures and guide precise geometric detailing. Current 3D generation methods predominantly focus on direct text to point cloud generation (Luo and Hu 2021; Nichol et al. 2022; Xiao et al. 2024), demonstrating feasibility but producing only latent representations with limited control over details. ProtoComp (Yu et al. 2024) generates coarse semantic prototypes using category text, yet fails to model local geometric relations. Category-level descriptions merely sketch rough outlines without part-level structural constraints.

Method

As illustrated in Figure 2 (a), the overall framework of our SGPRNet is presented. First, we utilize Multi-modal Large Language Models (MLLMs) to automatically generate the specific text of 3D shapes. Then, we design an Orthogonal Semantic Part Transfer (OSPT) module that learns transferable semantic relations between geometric parts. Finally, we develop a Semantic Geometric Relation-aware Transformer (SGRFormer) to form high consistency joint coarse point clouds utilizing these semantic relational features to progressively guide the point cloud refining process.

MLLMs-Assisted Text Corpus Generation

In this study, we leverage Qwen-vl (Bai et al. 2023) to automatically generate specific text that includes detailed text descriptions of the geometric relations between corresponding parts in rendered images from any perspective of the input point cloud, using 3D language command. We then use Electra (Clark et al. 2020) to extract text features F_{t} from these descriptions.

Multi-Scale Topology Encoder

Our goal is to generate coarse point clouds whose geometries align as closely as possible with the partial input. To achieve this, we propose a novel Multi-Scale Topology Encoder, which enhances the network's capability to infer geometric details in missing regions while capturing diverse local structural information of target shapes. Specifically, given the point coordinates P_{l-1} and corresponding point features F_{l-1}, we combine a set abstraction layer with a Point Transformer to model P_l and F_l at each level l. This process is described as follows:

$$(P_l, F_l) = \mathrm{PCT}(\mathrm{SA}(P_{l-1}, F_{l-1})), \tag{1}$$

where SA denotes the Set Abstract level (Qi et al. 2017), and PCT represents the Point Transformer (Zhao et al. 2021). By stacking multiple layers of SA and PCT, the final point coordinates P_{m}, the corresponding local point features F_{m}, and the global feature F_{g} are obtained.

Orthogonal Semantic Part Transfer Module

The encoder extracts a latent representation from the partial input. However, completing the point cloud requires a global semantic understanding for guidance, which the partial input alone cannot provide. Text vectors can effectively compensate for the missing global semantic information, especially the relations between geometric parts. Unlike previous approaches, we design a learnable Part Relation Dictionary (PRD) $D \in \mathbb{R}^{s \times d_p}$ and corresponding text descriptions of 3D shape, where s denotes the number of parts and d_p denotes the part dimension. This strategy not only helps generate more informative point cloud representations but also supplements point cloud representations with semantic information. The Orthogonal Semantic Part Transfer (OSPT) includes a Shared Feature Geometry (SFG) Module, a Part Semantic Enhancement (PSE) Module, and a Geometry Perception Mining (GPM) Module.

Shared Feature Geometry Module In this stage, we focus on learning collaborative alignment between geometric and text features to bridge the semantic gap between the text and point cloud modalities. We integrate a backbone network $\Phi(\cdot;\theta)$, analogous to a deconvolutional network (Zeiler et al. 2010), with several MLPs applied to F_{t}, F_{m}, and F_{g}. This process is expected to map F_{t} and F_{m} into a shared modal latent space with its parameters θ, generating coarse-grained geometric structure features $F_{\mathrm{pc}}^{\mathrm{c}}$ and coarse-grained part semantic features $F_{\mathrm{t}}^{\mathrm{c}}$, respectively. $F_{\mathrm{pc}}^{\mathrm{c}}$ and $F_{\mathrm{t}}^{\mathrm{c}}$ can be defined as follows:

$$F_{\mathrm{pc}}^{\mathrm{c}} = \Phi(\mathrm{MLP}(F_{\mathrm{m}}) \| (\mathrm{repeat}(F_{\mathrm{g}}); \theta)), \tag{2}$$

$$F_{\mathrm{t}}^{\mathrm{c}} = \Phi(\mathrm{MLP}(F_{\mathrm{t}}); \theta), \tag{3}$$

where $\|$ denotes channel-level concatenation.

Part Semantic Enhancement Module To enable human-like text-guided reasoning, we focus on learning geometric relations from a part relation dictionary and coarse semantic features$F_{\mathrm{t}}^{\mathrm{c}}$, refining these semantic features $F_{\mathrm{t}}^{\mathrm{c}}$ into fine-grained part semantic features $F_{\mathrm{t}}^{\mathrm{f}}$. Figure 2(b) shows the Part Semantic Enhancement (PSE) module, which integrates 3D geometric prior information into coarse features, ensuring semantic smooth guidance during the shape completion process. The design of PSE is simple and intuitive. Specifically, we calculate the similarity between PRD and coarse-grained part semantic features to identify the matching relations between part semantics and geometric structures. Subsequently, we aggregate these semantic relations with coarse-grained part semantic features, progressively refining the part semantic feature information to ultimately obtain fine-grained part semantic features. These fine-grained features effectively compensate for the lack of local detailed geometric structures, especially the connection relations between critical parts. In the Shared Feature Geometry module, its output $F_{\mathrm{t}}^{\mathrm{c}}$ serves as the query vector Q, while the part relation dictionary D serves as the key K and value V. Which is defined as follows:

$$Q = \psi_q(F_{\mathrm{t}}^{\mathrm{c}}), K = \psi_k(\varphi_N(D)), V = \psi_v(\varphi_N(D)), \tag{4}$$

$$F_{\mathrm{t}}^{\mathrm{f}} = \mathrm{softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}})V, \tag{5}$$

where φ_N is the normalization. ψ_q, ψ_k, and ψ_v are one-dimensional convolutional layers, $\frac{1}{\sqrt{d_k}}$ is the scaling factor, and d_k is the dimension of K. This process explicitly establishes a differentiable mapping of part semantic to geometric structures, allowing the model to generate physically plausible details based on linguistic descriptions despite heavy occlusion (e.g., "chair legs perpendicular to the chair surface" reconstructs "complete plane and perpendicular legs").

Similarly, as shown in Figure 2(b), we design the Geometric Perception Mining (GPM) module that refines coarse-grained geometric features $F_{\mathrm{pc}}^{\mathrm{c}}$ into fine-grained features $F_{\mathrm{pc}}^{\mathrm{f}}$ using shape information from the part relation dictionary D. This process enables cross-category knowledge transfer, meaning that features missing in the latent representations of some samples may be present in others. Additionally, we impose orthogonal constraints on the part relation dictionary D to enhance feature discriminability and capture the complementary relation between part semantics and geometric structures.

$$\mathcal{L}_{\mathrm{OD}} = \left\| \mathrm{I} - \frac{DD^T}{2\|DD^T\|} \right\|_2^2, \tag{6}$$

where I is the identity matrix.

Semantic Geometric Relation-aware Transformer

After obtaining the fine-grained geometric features, we use the part semantic relation information to guide the refinement process for local geometric relations, aiming to hierarchically fuse part semantic features across multi-scale representations. To achieve this, we propose a Semantic Geometric Relation-aware Transformer (SGRFormer). As shown in Figure 2(c), we design a geometric merging block comprising cross-attention and self-attention mechanisms to capture global dependencies while preserving local geometric correlations. Crucially, we employ a semantic-aware cross-attention layer at the terminal of the feature hierarchy to strengthen semantic guidance and ensure effective feature propagation, yielding fused features F_{f}, as illustrated below:

$$F_{\mathrm{f}} = \mathrm{M}(F_{\mathrm{pc}}^{\mathrm{f}}, F_{\mathrm{t}}^{\mathrm{f}}), \tag{7}$$

where M is a geometric merging block. We obtain the refined fused features F_{u} by adding fine-grained geometric features to the fused features F_{f}. Then F_{u} and feature channel augmented F_m^{a} by processing F_{m} with PCT and Multi-layer shared MLP are concatenated to form joint coarse point cloud features F_{j}. The joint coarse point cloud features can be defined as:

$$F_{\mathrm{u}} = \lambda F_{\mathrm{pc}}^{\mathrm{f}} + (1-\lambda) F_{\mathrm{f}}, \tag{8}$$

$$F_{\mathrm{j}} = F_{\mathrm{u}} \| F_m^{\mathrm{a}}, \tag{9}$$

where $\|$ denotes channel-wise concatenation. λ is the balancing coefficient between $F_{\mathrm{pc}}^{\mathrm{f}}$ and F_{f}. This multi-scale feature fusion significantly expands the representation space of the partial input point cloud while capturing local detailed structures between geometric parts.

Finally, we map joint coarse point cloud features F_{j} to coarse point clouds P_{c} using a single MLP, which are concatenated

with P_{m} to form the joint coarse point cloud P_{j}, providing an approximately complete shape prior. The joint coarse point cloud can be defined as:

$$P_{\mathrm{c}} = \mathrm{MLP}(F_{\mathrm{j}}), \tag{10}$$

$$P_{\mathrm{j}} = P_{\mathrm{m}} \oplus P_{\mathrm{c}}, \tag{11}$$

where $\oplus$ denotes point-level concatenation. Subsequently, we reconstruct detailed complete point cloud based on a coarse-to-fine pipeline with different upsampling ratios, which incorporates the joint coarse point cloud and its features.

Loss Function

In our implementation, we adopt the Chamfer Distance (CD) that computes the bidirectional average closest point distance to supervise the predicted complete point cloud against the ground truth. Given the ground truth G, we use P and J to present the predicted complete point cloud and the joint coarse point cloud, respectively. The CD can be defined as follows:

$$\mathcal{L}_0 = \frac{1}{|N_{\mathrm{P}}|} \sum_{p \in P} \min_{g \in G} \parallel p - g \parallel + \frac{1}{|N_{\mathrm{G}}|} \sum_{g \in G} \min_{p \in P} \parallel g - p \parallel, \tag{12}$$

$$\mathcal{L}_1 = \frac{1}{|N_{\mathrm{J}}|} \sum_{\mathrm{j} \in J} \min_{g \in G} \parallel j - g \parallel + \frac{1}{|N_{\mathrm{G}}|} \sum_{g \in G} \min_{\mathrm{j} \in J} \parallel g - j \parallel, \tag{13}$$

where N_P, N_{J}, and N_{G} denote the number of P, J and G points, respectively. Note that direct supervision of the joint coarse point cloud using ground truth aligns its distribution with the target shape while enhancing model self-consistency, thereby optimizing feature parameter learning. To sum up, the total training loss is expressed as follows:

$$\mathcal{L} = \alpha \mathcal{L}_0 + \beta \mathcal{L}_1 + \eta \mathcal{L}_{\mathrm{OD}}, \tag{14}$$

where $\mathcal{L}_{\mathrm{OD}}$ is the orthogonality constraints for the learnable part relation dictionary. α, β, and η are the balancing coefficient and set to 1 for all experiments.

Experiments

Dataset Introduction and Evaluation Metrics

In this section, to validate the effectiveness of specific text, we choose to generate samples based on the OmniObject3D dataset (Wu et al. 2023) as this dataset reflects more object categories in the real world and more occlusion situations from different perspectives. These categories, such as ShapeNet (Wu et al. 2015), ScanNet (Dai et al. 2017), and S3DIS (Armeni et al. 2017), are not available in this dataset. We then compare our method against existing point cloud completion methods on both seen and unseen categories. Additionally, we further validate the effectiveness of specific text using the widely used ShapeNet-ViPC dataset (Zhang et al. 2021). Finally, we provide ablation studies to more comprehensively demonstrate the effectiveness of our method.

Dataset To investigate the effectiveness of point cloud completion for specific text under varying occlusion conditions, we have created the point-text pair OmniObject3D-212/34 and Text-ViPC datasets based on the existing OmniObject3D and ShapeNet-ViPC datasets by incorporating specific text.

OmniObject3D-212/34 Dataset In the 212 seen categories, we randomly select 80% of the objects in each category as the training set (4,736 samples) and the remaining 20% as the test set (1,090 samples), denoted as the OmniObject3D/212 dataset. To systematically evaluate the model's generalization performance on unseen categories, we divide the 212 categories into 178 seen and 34 unseen categories, denoted as the OmniObject3D/34 dataset. It is worth noting that due to incomplete data in the OmniObject3D dataset, we remove some categories, resulting in an effective dataset with 212 categories. We then randomly select one of nine viewpoints with occlusion rates of 25% (light occlusion), 50% (moderate occlusion), and 75% (severe occlusion), and generate partial point clouds containing 12,288 (L), 8,192 (M), and 4,096 (S) points, respectively.

Text-ViPC Dataset Based on the commonly used ShapeNet-ViPC dataset, we have constructed the Text-ViPC dataset by incorporating specific text for rendered images. These images are selected from any of the 24 viewpoints within eight representative categories: table, chair, plane, cabinet, sofa, watercraft, car, and lamp. The training and testing sample settings are consistent with seen categories in XMFNet.

Evaluation Metric Following the previous works (Yuan et al. 2018; Tchapmi et al. 2019; Yu et al. 2021; Huang et al. 2020; Wei et al. 2025; Du et al. 2025), we compare our method's performance with existing works using ℓ_2 norm version of Chamfer Distance (ℓ_2-CD) and F-Score@1% (F1) as evaluation metrics.

Evaluation on the OmniObject3D/212 Dataset

For a fair comparison, we retrain all baseline models from our OmniObject3D/212 without incorporating text descriptions using their official open-source implementations and the optimal hyperparameters reported in their papers. As detailed in Table 1, which reports significant improvements in F1 and ℓ_2-CD for our method across all 212 categories, the results are further specified for four representative categories (bowl, cabinet, chair, and guitar) under moderate occlusion.

Evaluation on the OmniObject3D/34 Dataset

To demonstrate the generalization capability of our proposed method, we conduct comparative experiments on the OmniObject3D/34 dataset. Qualitative results confirm its significant effectiveness on unseen categories.

Our method surpasses all baselines on the 212 seen categories. Its true strength, however, is revealed on the 34 unseen categories (with$<$20 samples), where it exhibits exceptional robustness to severe occlusion. In fact, our method shows greater resilience on these unseen categories than

Method	Bowl	Cabinet	Chair	Guitar	L	M	S	Avg ℓ_2-CD/F1
PCN(3DV 2018)	2.070/0.674	5.024/0.426	51.328/0.285	8.203/0.415	7.402/0.519	12.475/0.433	18.209/0.342	12.695/0.431
AdaPointr(TPAMI 2023)	**2.040/0.680**	3.797/0.639	22.870/0.422	5.135/0.489	4.641/0.589	8.503/0.458	18.014/0.324	10.386/0.457
SnowflakeNet(TPAMI 2023)	8.753/0.371	6.124/0.652	40.039/0.441	13.832/0.433	6.330/0.611	15.293/0.409	27.565/0.244	16.396/0.421
SeedFormer(ECCV 2022)	4.638/0.572	**3.016/0.684**	30.855/**0.490**	11.096/0.547	4.209/**0.685**	9.439/0.521	17.987/**0.351**	10.545/**0.519**
SGPRNet(Ours)	2.115/0.658	3.059/0.673	**20.792**/0.480	**3.438/0.613**	**4.149**/0.664	**7.634/0.548**	**14.790**/0.345	**8.858/0.519**

Table 1: Mean Chamfer Distance and F1 score of different methods on the OmniObject3D/212 dataset(ℓ_2-CD $\times 10^3 \downarrow$ / F1↑).

Method	178 seen categories				34 unseen categories			
	L	M	S	Ave ℓ_2-CD/F1	L	M	S	Ave ℓ_2-CD/F1
PCN(3DV 2018)	8.135/0.507	11.976/0.431	18.514/0.326	12.875/0.421	16.094/0.427	30.145/0.359	37.272/0.277	27.837/0.354
AdaPointr(TPAMI 2023)	4.810/0.632	8.215/0.515	15.649/0.369	9.558/0.505	9.952/0.586	21.714/0.472	30.844/0.325	20.837/0.461
SnowflakeNet(TPAMI 2023)	6.831/0.624	14.539/0.408	27.501/0.233	16.290/0.422	11.535/0.615	28.070/0.399	38.726/0.242	26.110/0.419
SeedFormer(ECCV 2022)	4.450/**0.690**	8.937/**0.540**	16.258/**0.371**	9.882/**0.534**	9.346/**0.663**	21.897/**0.509**	27.986/**0.346**	19.743/**0.506**
SGPRNet(Ours)	**4.133**/0.652	**6.955**/0.520	**12.505**/0.356	**7.864**/0.509	**8.918**/0.612	**19.466**/0.480	**22.041**/0.326	**16.808**/0.473

Table 2: Mean Chamfer Distance and F1 score of different methods on the OmniObject3D/34 dataset (ℓ_2-CD$\times 10^3 \downarrow$ / F1↑).

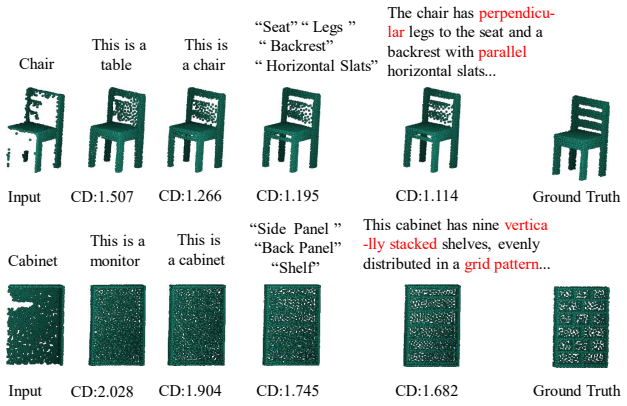

Figure 3: Visualized ablation study of two seen category objects on the Text-ViPC dataset. The first row and the fourth row display various specific texts, respectively.

on the seen ones. The proposed method demonstrates significant enhancements in F1 and ℓ_2-CD (see Table 2) on the OmniObject3D/34 dataset, which proves robust against varying occlusion levels and unseen categories.

Evaluation on the Text-ViPC Dataset

Our method, using only specific text and incomplete point clouds as input, substantially surpasses the state-of-the-art on both ℓ_2-CD and F1 metrics (see Table 3). Specifically, it reduces the average ℓ_2-CD loss by 0.101 and increases the average F1 by 0.026 compared to the previous best methods. While the quantitative statistics for EGIINet (Xu et al. 2024) are obtained directly, we utilize those for CDPNet (Du et al. 2024a) from CMNet (Du et al. 2025) to account for inconsistencies in its experimental setup.

As shown in Figure 4, ViPC tends to produce completed point clouds with blurred structures, particularly in areas like table legs and sofa armrests. While CSDN and CDPNet yield better qualitative results overall compared to ViPC, they still lack fine local geometric details, such as the junction between wings and fuselage or chair legs and seats. However, EGIINet fails to capture the geometric constraints at the junctions of chair back slats, such as their relative angles. In contrast, our method overcomes this limitation, leveraging text-generated joint coarse point clouds to constrain global structure and infer local details.

Ablation Studies

To validate our approach, this section conducts ablation studies to assess the contribution of different text inputs and key components to the final completion performance.

Contributions of Various Text Descriptions This section investigates how text descriptions affect the completion quality of 3D shapes. We observe that while basic category descriptions (e.g.,"This is a chair") improve overall performance, they still result in geometric ambiguities at critical part junctions. In contrast, detailed descriptions incorporating geometric structures and part relations (e.g.,"a cabinet with vertically stacked shelves") provide explicit semantic guidance to the network. As shown in Figure 3, the specific text likes "vertically stacked shelves and grid pattern" enables the network to generate well-defined divider structures for cabinet. Similarly, emphasizing "legs perpendicularly connected to the seat" significantly improves joint clarity between legs and seat surfaces for chair.

Contributions of Key Components To quantify the contribution of each component, we conduct ablation studies on the Text-ViPC dataset. As shown in Table 4, the results indicate that each component makes a substantial contribution to performance. Compared to the baseline model using a vanilla Transformer for point cloud completion, integrating

Method	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
ViPC(CVPR 2021)	3.308/0.591	1.760/0.803	4.558/0.451	3.183/0.512	2.476/0.529	2.867/0.706	4.481/0.434	4.990/0.594	2.197/0.730
CSDN(TVCG 2023)	2.570/0.695	1.251/0.862	3.670/0.548	2.977/0.560	2.835/0.669	2.554/0.761	3.240/0.557	2.575/0.729	1.742/0.782
XMFNet(NeurIPS 2022)	1.443/0.796	0.572/0.961	1.980/0.662	1.754/0.691	1.403/0.809	1.810/0.792	1.702/0.723	1.386/0.830	0.945/0.901
EGIINet(ECCV 2024)	1.211/0.836	0.534/0.969	1.921/0.693	1.655/0.723	1.204/0.847	0.776/0.919	1.552/0.756	1.227/0.857	0.802/0.927
CDPNet(AAAI 2024)	1.374/0.814	0.671/0.955	1.902/0.697	1.857/0.692	1.504/0.799	1.024/0.890	1.694/0.738	1.437/0.829	0.904/0.914
SGPRNet(Ours)	**1.110/0.862**	**0.499/0.978**	**1.639/0.749**	**1.597/0.746**	**1.096/0.876**	**0.725/0.927**	**1.425/0.796**	**1.181/0.875**	**0.717/0.946**

Table 3: Mean Chamfer Distance and F1 score of different methods on the Text-ViPC dataset(ℓ_2-CD $\times 10^3 \downarrow$ / F1$\uparrow$).

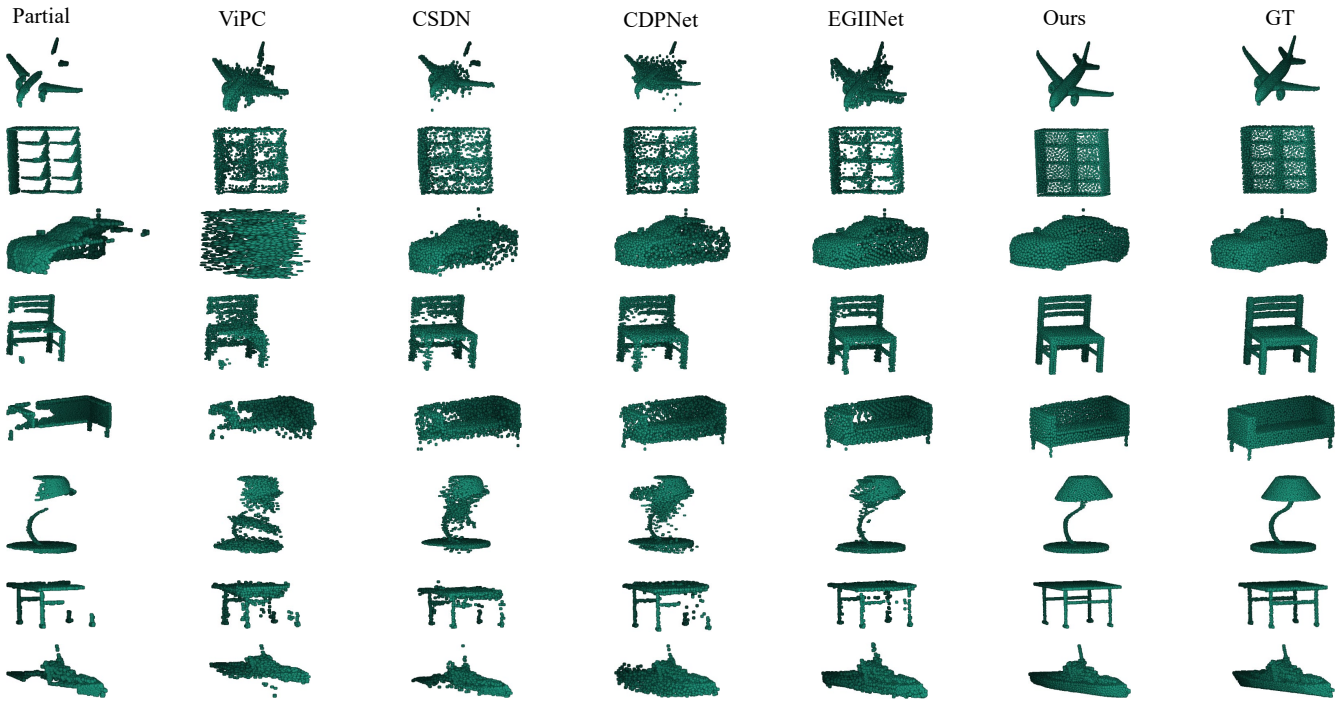

Figure 4: Visualization results on the Text-ViPC dataset.

PSE	GPM	SGRFormer	Ave ℓ_2-CD	Ave-F1
			1.605	0.726
✓			1.332	0.819
	✓		1.458	0.793
		✓	1.529	0.735
✓	✓		1.186	0.829
✓		✓	1.172	0.836
	✓	✓	1.389	0.817
✓	✓	✓	1.110	0.862

Table 4: Ablation results of different components on the Text-ViPC dataset.

the PSE reduces the average ℓ_2-CD by 0.273 and improves the average F1 score by 0.093. Most notably, the PSE yields the most significant improvements. This results not only validate the effectiveness of each component but also highlight the critical role of PSE in enhancing the representational capabilities of the upsampling process.

Conclusion

In this work, we present SGPRNet, a novel point cloud completion framework designed for the real world scenarios. Unlike existing methods, our approach introduces a description of the geometric relations between parts to enhance the understanding of geometric structures. In addition, we introduce MLLMs to automatically generate detailed text descriptions of the geometric relations between parts, which significantly improves the framework's ability to capture semantic geometric information and represent complex 3D spatial structures. While achieving promising results, we also realize that high-quality text descriptions require the design of sophisticated 3D language command for MLLMs, which is costly and too subjective. In future work, we will investigate the controlled point cloud completion framework, which may reduce the reliance on designing 3D language command for MLLMs while maintaining high fidelity and consistency of reconstructed geometric structures.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U21A20473, 62376142) 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. *Advances in Neural Information Processing Systems*, 35: 37349–37362.

Armeni, I.; Sax, S.; Zamir, A. R.; and Savarese, S. 2017. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Cao, X.; Zhou, T.; Ma, Y.; Ye, W.; Cui, C.; Tang, K.; Cao, Z.; Liang, K.; Wang, Z.; Rehg, J. M.; et al. 2024. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21819–21830.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cui, R.; Qiu, S.; Anwar, S.; Liu, J.; Xing, C.; Zhang, J.; and Barnes, N. 2023. P2c: Self-supervised point cloud completion from single partial clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14351–14360.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Dai, A.; Ruizhongtai Qi, C.; and Nießner, M. 2017. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5868–5877.

Du, Z.; Dou, J.; Liu, Z.; Wei, J.; Wang, G.; Xie, N.; and Yang, Y. 2024a. CDPNet: cross-modal dual phases network for point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1635–1643.

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

Du, Z.; Liu, Z.; Wang, G.; Wei, J.; Yussif, S. B.; Wang, Z.; Xie, N.; and Yang, Y. 2025. CMNet: Cross-Modal Coarse-to-Fine Network for Point Cloud Completion Based on Patches. *IEEE Transactions on Circuits and Systems for Video Technology*.

Duan, F.; Yu, J.; and Chen, L. 2024. T-CorresNet: template guided 3D point cloud completion with correspondence pooling query generation strategy. In *European Conference on Computer Vision*, 90–106. Springer.

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.

Han, X.-F.; Laga, H.; and Bennamoun, M. 2019. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(5): 1578–1604.

Hou, J.; Dai, A.; and Nießner, M. 2019. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4421–4430.

Huang, Z.; Yu, Y.; Xu, J.; Ni, F.; and Le, X. 2020. Pf-net: Point fractal network for 3d point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7662–7670.

Li, S.; Gao, P.; Tan, X.; and Wei, M. 2023. 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*, 9466–9475.

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*, 45(12): 14975–14989.

Liu, M.; Sheng, L.; Yang, S.; Shao, J.; and Hu, S.-M. 2020. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11596–11603.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2837–2845.

Lyu, Z.; Kong, Z.; Xu, X.; Pan, L.; and Lin, D. 2021. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30: 5105–5114.

Stutz, D.; and Geiger, A. 2018. Learning 3d shape completion from laser scan data with weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1955–1964.

Tchapmi, L. P.; Kosaraju, V.; Rezatofighi, H.; Reid, I.; and Savarese, S. 2019. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 383–392.

Varley, J.; DeChant, C.; Richardson, A.; Ruales, J.; and Allen, P. 2017. Shape completion enabled robotic grasping. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2442–2447.

Wang, W.; Xie, J.; Hu, C.; Zou, H.; Fan, J.; Tong, W.; Wen, Y.; Wu, S.; Deng, H.; Li, Z.; et al. 2023. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*.

Wang, X.; Ang Jr, M. H.; and Lee, G. H. 2020. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 790–799.

Wei, G.; Feng, Y.; Ma, L.; Wang, C.; Zhou, Y.; and Li, C. 2025. Pcdreamer: Point cloud completion through multi-view diffusion priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27243–27253.

Wu, T.; Zhang, J.; Fu, X.; Wang, Y.; Ren, J.; Pan, L.; Wu, W.; Yang, L.; Wang, J.; Qian, C.; et al. 2023. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 803–814.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Xiang, P.; Wen, X.; Liu, Y.-S.; Cao, Y.-P.; Wan, P.; Zheng, W.; and Han, Z. 2022. Snowflake point deconvolution for point cloud completion and generation with skip-transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 6320–6338.

Xiao, H.; Kang, W.; Li, Y.; and Xu, H. 2024. Text-Free Controllable 3-D Point Cloud Generation. *IEEE Transactions on Instrumentation and Measurement*, 73: 1–12.

Xie, H.; Yao, H.; Zhou, S.; Mao, J.; Zhang, S.; and Sun, W. 2020. Grnet: Gridding residual network for dense point cloud completion. In *European conference on computer vision*, 365–381. Springer.

Xiong, Y.; Ma, W.-C.; Wang, J.; and Urtasun, R. 2023. Learning compact representations for lidar completion and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1074–1083.

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 *European Conference on Computer Vision*, 414–432. Springer.

Ying, H.; Shao, T.; Wang, H.; Yang, Y.; and Zhou, K. 2023. Adaptive local basis functions for shape completion. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

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*, 12498–12507.

Yu, X.; Wang, Y.; Zhou, J.; and Lu, J. 2024. ProtoComp: Diverse Point Cloud Completion with Controllable Prototype. In *European Conference on Computer Vision*, 270–286. Springer.

Yuan, W.; Khot, T.; Held, D.; Mertz, C.; and Hebert, M. 2018. Pcn: Point completion network. In *2018 international conference on 3D vision (3DV)*, 728–737. IEEE.

Zeiler, M. D.; Krishnan, D.; Taylor, G. W.; and Fergus, R. 2010. Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, 2528–2535. IEEE.

Zhang, W.; Yan, Q.; and Xiao, C. 2020. Detail preserved point cloud completion via separated feature aggregation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 512–528. Springer.

Zhang, X.; Feng, Y.; Li, S.; Zou, C.; Wan, H.; Zhao, X.; Guo, Y.; and Gao, Y. 2021. View-guided point cloud completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15890–15899.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16259–16268.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.

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 *European conference on computer vision*, 416–432. Springer.

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

Zhu, Z.; Nan, L.; Xie, H.; Chen, H.; Wang, J.; Wei, M.; and Qin, J. 2023b. Csdn: Cross-modal shape-transfer dual-refinement network for point cloud completion. *IEEE Transactions on Visualization and Computer Graphics*, 30(7): 3545–3563.