

Position-Aware Guided Point Cloud Completion with CLIP Model

Feng Zhou¹, Qi Zhang¹, Ju Dai^{2*}, Lei Li^{3,4}, Qing Fan⁵, Junliang Xing⁶

¹North China University of Technology, Beijing, China

²Peng Cheng Laboratory, Shenzhen, China

³University of Copenhagen, Copenhagen, Denmark

⁴University of Washington, Washington, USA

⁵Skywork AI, Beijing, China

⁶Tsinghua University, Beijing, China

zhoufeng@ncut.edu.cn, zhangqi@mail.ncut.edu.cn, daij@pcl.ac.cn

lilei@di.ku.dk, qing.fan@kunlun-inc.com, jlxing@tsinghua.edu.cn

Abstract

Point cloud completion aims to recover partial geometric and topological shapes caused by equipment defects or limited viewpoints. Current methods either solely rely on the 3D coordinates of the point cloud to complete it or incorporate additional images with well-calibrated intrinsic parameters to guide the geometric estimation of the missing parts. Although these methods have achieved excellent performance by directly predicting the location of complete points, the extracted features lack fine-grained information regarding the location of the missing area. To address this issue, we propose a rapid and efficient method to expand an unimodal framework into a multimodal framework. This approach incorporates a position-aware module designed to enhance the spatial information of the missing parts through a weighted map learning mechanism. In addition, we establish a Point-Text-Image triplet corpus PCI-TI and MVP-TI based on the existing unimodal point cloud completion dataset and use the pre-trained vision-language model CLIP to provide richer detail information for 3D shapes, thereby enhancing performance. Extensive quantitative and qualitative experiments demonstrate that our method outperforms state-of-the-art point cloud completion methods.

Introduction

Point clouds are a fundamental data structure across various domains, including autonomous driving (Chen, Dai, and Ding 2022; Mao et al. 2023; Cheng and Li 2024), robotics (Christen et al. 2023) and others (Oehmcke et al. 2022; Han et al. 2020; Nguyen et al. 2016; Oehmcke et al. 2024). However, in real-world scenarios, challenges such as object occlusions, variability in surface material reflectivity, and limitations in sensor resolution and field of view frequently result in the acquisition of incomplete point cloud data. These deficiencies impede the effectiveness of downstream applications, underscoring the critical need for point cloud completion. Therefore, point cloud completion is indispensable, and it has attracted more research interest in recent years.

Leveraging large-scale point cloud datasets (Yuan et al. 2018; Yu et al. 2021), many learning-based point cloud completion methods have achieved excellent results (Zhang,

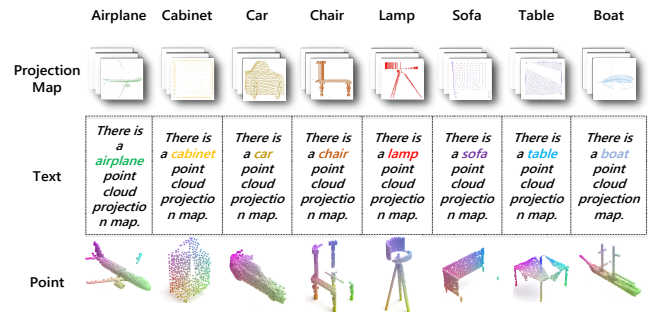


Figure 1: An exemplary instance from our PCN-TI dataset. The PCN-TI dataset comprises many triplets (Point-Text-Image) consisting of projection images, readily available textual descriptions, and incomplete point clouds.

Yan, and Xiao 2020; Zhang et al. 2022c; Zhou et al. 2022; Zhang et al. 2022a; Li 2023). However, despite these advancements, there remains significant potential for further improvement.

The first aspect is that previous studies have identified that the missing parts either share similar structural information with the incomplete point cloud or are consistent with some existing structures. Consequently, numerous endeavors learn the prior shape information of the incomplete point cloud or suitable geometric patterns to tackle the point cloud completion challenge. For example, (Yuan et al. 2018) proposes an encoder-decoder network to learn global shape information that directly encodes partial shapes into a global feature and then decodes it into complete shapes. (Tang et al. 2022) thinks that despite the absence of some point cloud data, the incomplete point cloud still maintains the principal skeletal structure. The authors propose a completion strategy that follows the "Keypoints-Skeleton-Shape" paradigm. By identifying and aligning key points, and then leveraging these key points with geometric priors, they introduce an innovative structure known as the surface skeleton, thereby acquiring comprehensive topological data and enhancing the information of local details.

The second aspect is that although the above methods can

*The corresponding author

obtain plausible results, the incomplete point clouds may lack critical semantic parts, making it difficult for point-based networks to recognize reasonable global shapes and accurately locate missing regions. Color images can provide rich texture and color information to assist in recovering the missing geometric structure of the point cloud data (Zhu et al. 2023b). However, pairing these images with point cloud data usually requires precise camera calibration and a complex spatial alignment process, which is very challenging in practical applications.

The third aspect is that while images can substantially aid in reconstructing the missing geometric structures of point cloud data, the unordered nature of point clouds leads to a significant challenge in precisely correlating the image information with the missing point cloud data. With the advent of visual-language pre-training (VLP) models such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021) and CoCa (Yu et al. 2022), an attractive prevalence has emerged: *Is it feasible to leverage VLP model to locate the missing points in point clouds?* In the field of 2D images, there are already many works have been conducted on position-guided text prompts. For example, (Wang et al. 2023) introduces a method to improve the spatial understanding of VLP models by using strategically designed text prompts, which leads to better performance on vision-language tasks that require precise localization and reasoning about objects in images. (Zhang et al. 2021a) employ a Faster-RCNN model to extract salient region features and model the location information using bounding boxes. However, images rendered or projected from point clouds usually lack the necessary texture information, and since most images contain only a single object, it becomes challenging to accurately find the missing object parts in the image using the above techniques.

In this paper, we propose a method that can quickly and efficiently expand an unimodal framework into a multimodal framework to address the aforementioned issues. It can simultaneously fuse the extra visual and textual information derived from the point cloud and provide precise local-global positional information for the model to ascertain where the incomplete point cloud should be completed, as shown in Figure 2. First, unlike previous multimodal point cloud completion paradigms that focus on the fusion of point images, we introduce additional text descriptions into our model to enhance its capabilities. Drawing inspiration from (Song et al. 2023), which introduces supplementary text descriptions for datasets comprising point-image pairs, our method differs from (Song et al. 2023) in that the textual descriptions we generate do not rely on LLMs, nor do they require additional images. Our method depends solely on the unimodal point cloud completion dataset itself, avoiding complex pre-processing steps, and is conducive to the rapid and efficient creation of a Point-Text-Image corpus, as shown in Figure 1. Considering that it is difficult for text descriptions to accurately locate the missing position information of incomplete point clouds, this can result in reduced performance of point cloud completion. Therefore, inspired by the work (Wang et al. 2023), which divides the image into multiple blocks, performs relevant operations on each block, and demonstrates great generalization

across a variety of tasks. We also divide the projection image into different blocks and adaptively learn weight information for each block to obtain local position information to ascertain if the current block corresponds to the projection location of the missing part. Then, we inpaint the incomplete projection maps to obtain the global position information. The obtained local-global information facilitates the positioning of the incomplete point cloud. Moreover, to validate the effectiveness of the multimodal approach proposed in this paper, we conduct extensive experiments on the PCN and MVP datasets. Additionally, building upon the aforementioned method of extending from an unimodal to a multimodal framework, we introduce two multimodal datasets, PCN-TI and MVP-TI. These new datasets transform the initial unimodal datasets, which solely comprise point cloud data, into comprehensive multimodal triplet datasets encompassing point clouds, text, and images.

The main contributions can be summarized as follows:

- We propose a method for rapidly and efficiently transforming an unimodal point cloud completion framework into a multimodal point cloud completion framework.
- We design a position-aware module to learn the location information of the missing parts of the point cloud, enabling the network to be more targeted during the completion process.
- For each point cloud, we introduce a paired textual description and projection maps corpus, which can provide richer descriptive details, and we have proposed two extended datasets, PCN-TI and MVP-TI. Extensive experiments demonstrate the superior performance of our method against previous state-of-the-art methods.

Related Work

Point-based Point Cloud Completion

With the advancements in network architectures designed for point clouds, especially the emergence of PointNet/PointNet++ (Qi et al. 2017a,b), point-based approaches have become the mainstream solution for point cloud completion, and remarkable progress has been achieved. PCN (Yuan et al. 2018) employs a deep neural network designed specifically for processing and completing point clouds. It adopts a coarse-to-fine architecture to generate a rough approximation of the missing parts first and then refine the details to achieve a more accurate completion. Based on the encoder-decoder architecture, many works (Cai et al. 2024; Wen et al. 2021, 2022; Xiang et al. 2021) obtain plausible performance. For example, SnowflakeNet (Xiang et al. 2021) interprets point cloud completion as an explicit and structured generation of local patterns and introduces a novel type of skip transformer to learn the split patterns in the Snowflake Point Decomposition (SPD). This module learns the shape context and spatial relationships between child and parent points, generating locally structured and compact point arrangements, and captures the structural features of 3D surfaces within local patches, significantly enhancing the performance of 3D shape completion. More recently,

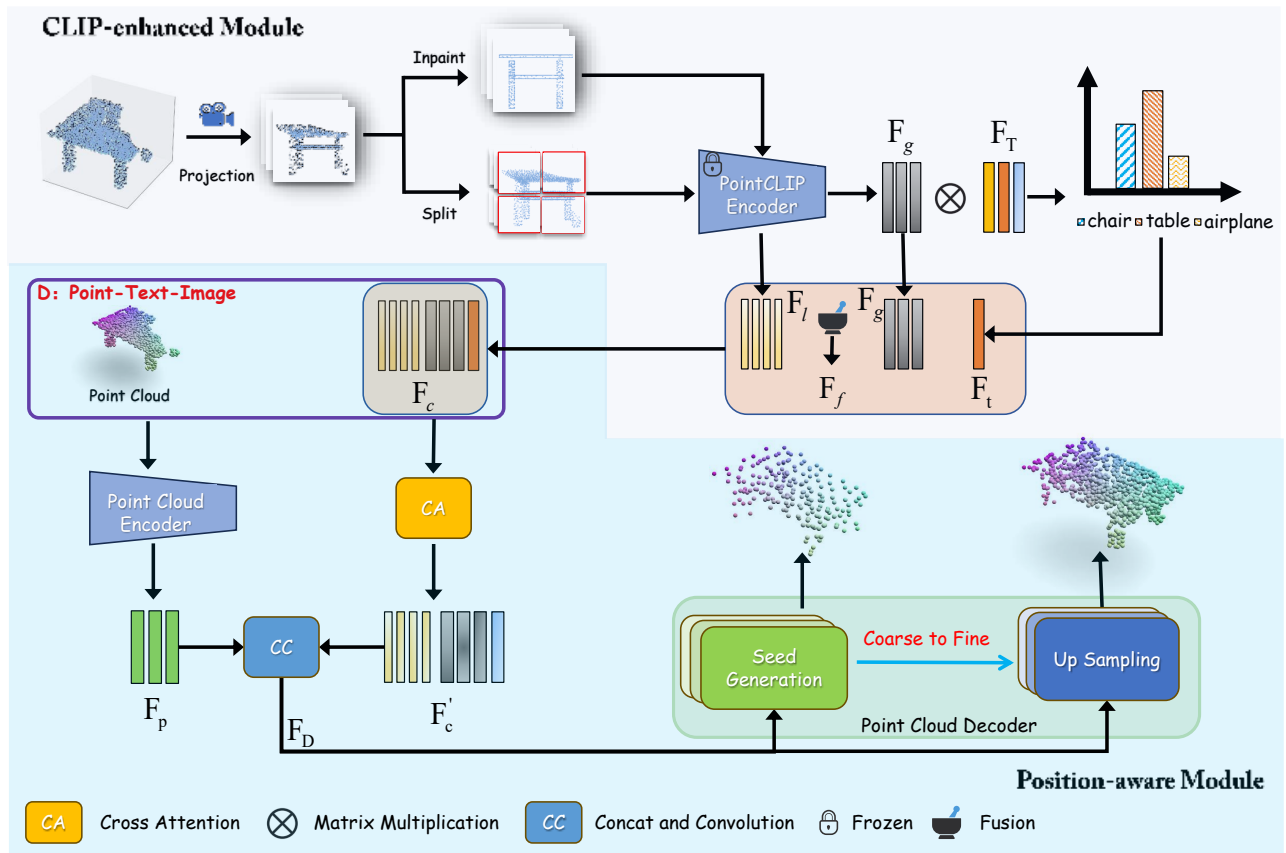


Figure 2: The overall architecture of our method consists of two main parts: the CLIP-enhanced module and the Position-aware module. F_g , F_l , F_T , and F_t in the CLIP-enhanced module denote the global-scale feature, the local-scale feature, the text feature from CLIP, and the processed text feature of F_T , respectively. F_c , F'_c , F_p , and F_D denote the CLIP feature, processed CLIP feature, point cloud feature, and fusion feature fed into the decoder, respectively.

AdaPoinTr (Yu et al. 2023) and PoinTr (Yu et al. 2021) convert the point cloud into a series of point proxies and employ a Transformer-based Encoder-Decoder for the generation of missing point clouds. Although these methods have achieved attractive results, due to the inherent limitations of singular modality, they cannot overcome the inherent ambiguity of incomplete data.

Multimodal Point Cloud Completion

Due to the incompleteness of the collected point clouds, the amount of information is limited, leading to significant uncertainty when inferring missing points. Additionally, point clouds are unstructured data with inherent sparsity, making it difficult to determine whether the blank 3D space is due to inherent sparsity or incompleteness, resulting in a lack of model interpretability. Many methods leverage other modalities that contain the necessary structural or semantic information of the missing parts of the given shape to complement the missing information of the point cloud, such as (Zhang et al. 2021b; Aiello, Valsesia, and Magli 2022; Kasten, Rahamim, and Chechik 2024; Song et al. 2023). ViPC (Zhang et al. 2021b) proposes to enhance the quality of complete shapes by utilizing the complementary modal informa-

tion, integrating the local information provided by the missing point cloud with the global structural information provided by a single view to accomplish the task of point cloud completion through multi-modal fusion. (Zhu et al. 2023b) employs a coarse-to-fine completion paradigm, addressing the challenge of cross-modal data fusion through two key modules: shape fusion and dual refinement. The shape fusion module utilizes IPAdaIN to guide the geometric generation of missing areas. Subsequently, the dual refinement module enhances the accuracy of details by adjusting the positions of the generated points. Multimodal methods have achieved encouraging results, and our proposed method also adopts the paradigm of multimodal fusion. However, most methods expand the content of multimodal information by adding text descriptions or additional images or by combining these approaches. However, text descriptions struggle to specify the exact location of the missing parts in incomplete point clouds, and the added image requires complex alignment to be utilizable. In this paper, we adopt a self-structure-based approach on the point cloud to harness additional image information while leveraging the detailed information provided by text, combining both to obtain location information to solve this problem.

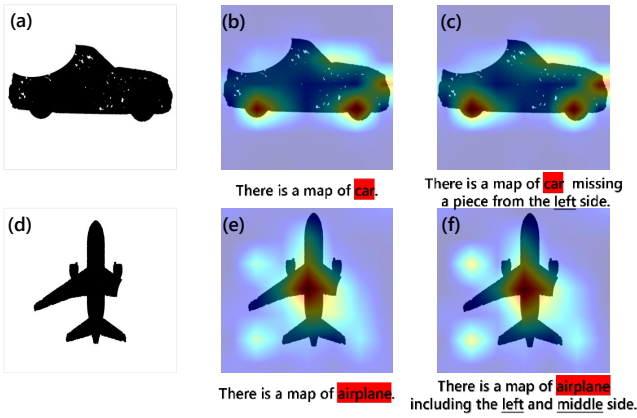


Figure 3: Our starting point for the position-aware module is as follows: (a) a projection image projected from an incomplete car point cloud, (b) the relevancy between the text "This is a map of car" and the projection map; (c) the relevancy between the text "There is a map of car missing a piece from the left side" and the projection map. (d) a projection map derived from an incomplete airplane point cloud. To illustrate the potential limitations of texts in capturing missing parts, we provide (e) and (f) for fair comparisons.

Methodology

Overview

The overall architecture of our method is shown in Figure 2. Given an incomplete and sparse point cloud $P \in \mathbb{R}^{N_p \times 3}$, the category information of P is expanded to generate a sentence T , followed by projecting P onto the coordinate axes to obtain six projection images $I = \{I_1, I_2, \dots, I_6\}$. The point cloud P , text sentence T , and the six projection images I are then paired as a pair of data $D = \{P, T, I\}$. This preliminary preprocessing data D serves as input for the network. Our goal is to infer the missing shapes of \hat{P} and produce a complete and dense point cloud $O \in \mathbb{R}^{N_o \times 3}$.

CLIP-enhanced Module

To explore the effectiveness of multimodality in point cloud completion, we provide a straightforward and practical approach to expanding existing datasets. By leveraging our proposed CLIP-enhanced module, it is possible to rapidly construct a new multimodal dataset based on current point cloud completion datasets, which can then be applied to existing CLIP methods to boost performance. Since the CLIP-enhanced module is not tailored to any specific dataset, it can be applied to common point cloud completion datasets. Here, we use the PCN dataset as an example for illustration.

We construct the text-image corpus called PCN-TI based on the PCN dataset. PCN-TI contains 30,974 pairs of text and images, forming triples along with the data in PCN, consisting of a point cloud, a text description, and a set of projection images. Some example triples are visualized in Figure 1. The text generation is implemented on a single NVIDIA RTX 4090. We generate the text description of each point cloud as follows: *There is $\{*\}$ point cloud pro-*

jection map, where $\{*\}$ denotes the category of the input point cloud. Regarding the projection map, similar to (Zhang et al. 2022b), we perform orthogonal projections for three-dimensional coordinates (x, y, z) of the input incomplete point cloud onto six faces to obtain richer two-dimensional projection information. By assigning the depth values of each point to discrete coordinates, we form an initial projection map. To further enhance the quality of the projection map, we conduct normalization operations on each map.

Our text corpus generation in the CLIP-enhanced module is similar to (Song et al. 2023). It enriches the expressiveness of the original dataset by incorporating textual information. Unlike (Song et al. 2023), our method of obtaining text does not require large language models (LLMs) or complex pre-processing procedures. The text generation of (Song et al. 2023) leverages LLMs to provide very fine-grained textual descriptions. For instance, for a Lamp, (Song et al. 2023) offers a description like "This is a *spiegel-symmetric lamp. The lamp has a spherical base and a circular curved shade. The lamp has one light bulb and a long stem.*" The length of the generated text descriptions in their corpus ranges from 50 to 58. In contrast, our method can achieve the goal simply through keyword substitution without relying on LLMs for complex content generation. For example, the text description in our method about lamp category is: *There is lamp point cloud projection map.*

Subsequently, the generated text T , along with the six projection images, are fed into the CLIP model. Since CLIP only accepts one text-image pair $\{T, I\}$, we utilize six identical CLIP models to process each projection image $\{I_1, I_2, \dots, I_6\}$ separately to get F_t, F_g , and then concatenated the obtained features F_l together to form F_c , the detail please refer to the CLIP-enhanced module in Figure 2.

Position-aware Module

While the CLIP-enhanced module provides substantial multimodal dataset support for network training, accurately identifying the missing location information in incomplete point clouds remains challenging for the network during learning. To address this issue, the initial trying is to integrate the positional information of the incomplete point cloud into the generated text description. We explore several prompts that encode positional information, for example:

- The projection of this $\{*\}$ is missing a piece in the Top Right corner.
- The top left and bottom right of the $\{*\}$ is missing.
- The $\{*\}$ is missing the top left and bottom right.

where $\{*\}$ denotes the class, and we adopt several other similar prompts. However, we find it challenging to accurately obtain positional information on the image through text descriptions alone. To verify this hypothesis, we utilize (Chefer, Gur, and Wolf 2021) to visualize the attention regions of both text and images. We conduct two sets of comparative experiments. In the first, an incomplete point cloud is presented alongside a description indicating the location of the missing parts, as depicted in Figure 3. By comparing Figure 3(b) and (c), we observe that the positional information given in the text, "missing a piece from the left side,"

does not align well with the corresponding location in the image. To avoid the issue of misalignment between text descriptions and images due to the missing parts not being visible in the image, we present a comparative experiment in Figure 3 (e) and (f), where the text description includes positional information about existing parts in the image: "including the left and middle side." This also does not achieve a satisfactory alignment with the relevant image areas.

To effectively address this issue, inspired by the work of (Wang et al. 2023), we divide the obtained projection map into equal-size non-overlapping blocks to get good performance. The details are as follows. First, we feed the obtained projection image into the image encoder module of CLIP. The ViT-16 (Dosovitskiy et al. 2021) model is leveraged as our image encoder. The obtained projection images are segmented into 2×2 blocks in the experiments, and parameter learning is conducted for these blocks. We randomly select one block in each training iteration to learn its parameters while setting the others to a default value of 1. In testing, all well-trained parameters are loaded, and these 24 parameters are subjected to cross-attention operations with the previously acquired image features to obtain the local-scale feature F_l that captures local-scale missing part information.

Additionally, in our experiments, inspired by (Song et al. 2023), we find that although the ShapeNet-ViPC dataset does not incorporate text descriptions, it introduces complete rendering images. (Zhang et al. 2021b; Zhu et al. 2023b) have already demonstrated that this additional complete rendering image provides good guidance information (such as view-guided information and image-guided information) for the point cloud, significantly enhancing the completion performance. Therefore, in our experiments, we utilize (Nazeri et al. 2019) to inpainting each projection image and then feed it into the ViT-16 to acquire a global-scale missing location feature F_g .

Then, to fuse F_l and F_g , we adopt feature fusion layers similar to (Zhu et al. 2023a) and output the final image feature F_f . The local-scale feature F_l is first transformed to query, key, and value tokens via linear projection and the guidance of the global-scale feature F_g . Then, the output feature F_f is obtained after two matrix multiplication.

To this end, we feed the input point cloud P into the point cloud encoder to obtain feature F_p . The previously obtained CLIP features F_C are fed into the cross-attention module, resulting in transformed CLIP features F'_C . We concatenate F_p and F'_C and feed the concatenated result into a 1×1 convolutional layer to obtain fusion feature F_D , which serves as the input of the point cloud decoder to generate the complete point cloud. Note that the architectures of our point cloud encoder and decoder depend on the baseline we choose.

Experiments

Datasets and Evaluation Metrics

PCN: The PCN dataset (Yuan et al. 2018) is a subset of ShapeNet dataset (Chang et al. 2015), which has 8 categories and contains 30,974 pairs of partial and complete point clouds. Incomplete point clouds are generated by projecting complete shapes onto eight partial views. For each

Methods	Ave ↓	Air	Cab	Car	Cha	Lam	Sof	Tab	Boat
FoldingNet	14.31	9.49	15.80	12.61	15.55	16.41	15.97	13.65	14.99
TopNet	12.15	7.61	13.31	10.90	13.82	14.44	14.78	11.22	11.12
PCN	9.64	5.50	22.70	10.63	8.70	11.00	11.34	11.68	8.59
GRNet	8.83	6.45	10.37	9.45	9.41	7.96	10.51	8.44	8.04
PoinTr	8.38	4.75	10.47	8.68	9.39	7.75	10.93	7.78	7.29
SnowflakeNet	7.21	4.29	9.16	8.08	7.89	6.07	9.23	6.55	6.40
FBNet	6.94	3.99	9.05	7.90	7.38	5.82	8.85	6.35	6.18
ProxyFormer	6.77	4.01	9.01	7.88	7.11	5.35	8.77	6.03	5.98
SeedFormer	6.74	3.85	9.05	8.06	7.06	5.21	8.85	6.05	5.85
AnchorFormer	6.59	3.70	8.94	7.57	7.05	5.21	8.40	6.03	5.81
AdaPoinTr	6.53	3.68	8.82	7.47	6.85	5.47	8.35	5.80	5.76
ODGNet	6.50	3.77	8.77	7.56	6.84	5.09	8.47	5.84	5.66
CRA-PCN	6.39	3.59	8.70	7.50	6.70	5.06	8.24	5.72	5.64
Ours	6.34	3.64	8.57	7.45	6.60	4.91	8.21	5.72	5.58

Table 1: Results on PCN dataset in terms of L1 CD $\times 10^3$ (lower is better) with SOTA methods.

complete shape, 16,384 points are uniformly sampled from the surface of the CAD model. The dataset is partitioned similarly to PCN to ensure a fair comparison of our method with other methods. Concurrently, following prior work, the sampled points are down-sampled to a standardized size of 2,048 points for training purposes.

MVP: The MVP dataset consists of 16 categories of high-quality pairs of partial and complete point clouds for training and testing. Each point cloud is captured from 26 uniformly distributed camera poses, offering a rich dataset for multi-view analysis and learning. Eight of the 16 categories (airplane, cabinet, car, chair, lamp, sofa, table, and watercraft) are the same as (Yuan et al. 2018), and another eight categories (bed, bench, bookshelf, bus, guitar, motorbike, pistol, and skateboard) are also included.

Proposed PCN-TI and MVP-TI: To explore the effectiveness of text descriptions and projection images in our method, we build a Point-Text-Image corpus called PCN-TI and MVP-TI based on the above two datasets. PCN-TI contains 30,974 triples, which are divided into 8 categories, the same as the PCN dataset. MVP-TI consists of high-quality pairs of partial and complete point clouds of 16 categories, the same as the MVP dataset.

Baselines. We evaluate four variants of pre-training models, including SnowflakeNet (Xiang et al. 2021), PoinTr (Yu et al. 2021), AdaPoinTr (Yu et al. 2023), and CRA-PCN (Rong et al. 2024), for their superior performance. If not explicitly specified, our default baseline is CRA-PCN model.

Evaluation metrics. One of the challenges in point cloud completion is the comparison with the ground truth. In this paper, we follow the existing work (Yuan et al. 2018; Xiang et al. 2021; Rong et al. 2024) to use the L1 CD, L2 CD, $F1$ -Score@1%, Fidelity, and MMD as the evaluation metrics. CD (Chamfer Distance) calculates the average closest point distance between the output point cloud O and the ground truth point cloud. Fidelity denotes the average distance from each point x in input P to its nearest neighbor in the output O . MMD (Minimal Matching Distance) measures how much the output resembles a typical car.

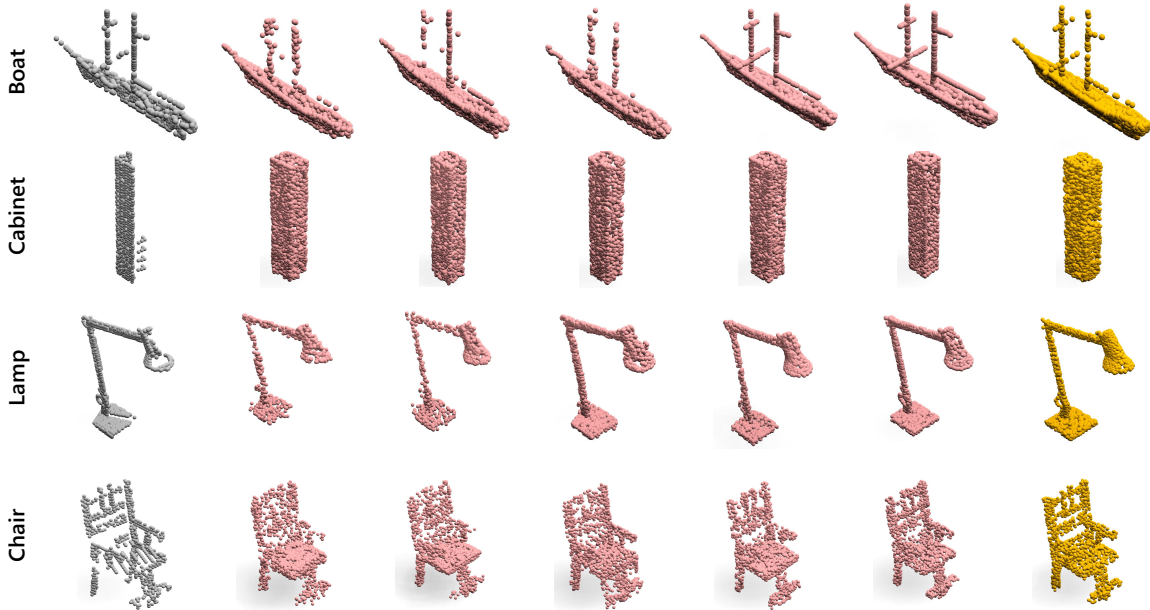


Figure 4: Point cloud completion results on PCN dataset. From left to right: partial input, results of GRNet, PoinTr, SnowflakeNet, AdaPoinTr, ours and ground truth. Best viewed in color and zoom in.

Results on Existing Benchmarks

Results on PCN dataset. We evaluate our method with many state-of-the-art (SOTA) methods, including FoldingNet (Yang et al. 2018), TopNet (Tchapmi et al. 2019), PCN (Yuan et al. 2018), GRNet (Xie et al. 2020), PoinTr (Yu et al. 2021), SnowflakeNet (Xiang et al. 2021), FBNet (Yan et al. 2022), ProxyFormer (Li et al. 2023), SeedFormer (Zhou et al. 2022), AnchorFormer (Chen et al. 2023), AdaPoinTr (Yu et al. 2023), ODGNet (Cai et al. 2024), CRA-PCN (Rong et al. 2024). The details are shown in Table 1. The quantitative results demonstrate that our method achieves the best performance across the CD metric. Specifically, our method achieves the best performance on 7/8 categories, which proves its robust generalization ability for completing shapes across different categories. Among the compared methods, PCN and SnowflakeNet are typical point cloud completion models that generate complete point clouds based on an encoder-decoder diagram via a max-pooling operation. Our method can achieve better results based on this encoder-decoder diagram. Meanwhile, PoinTr and AdaPoinTr reformulate point cloud completions as a set-to-set translation problem. Compared with these methods, our method can still perform better. Therefore, the improvements should be credited to our CLIP-enhanced module and the position-aware module, which help to overcome the limitations of traditional unimodal methods while introducing more precise information about missing positions, enabling the generation of points with greater accuracy.

Like the practice in the PCN dataset and many other works, we visually compare our method with SOTAs in Figure 4. The visual results show that our method can predict the complete point clouds with much better shape quality.

Method	Points	Ave ↓	$F1$ -Score@1% ↑
PCN	2048	9.77	0.321
TopNet	2048	10.11	0.308
ECG	2048	7.25	0.434
CRN	2048	6.64	0.476
VRCNet	2048	5.96	0.499
PoinTr	2048	6.15	0.456
CRA-PCN	2048	5.33	0.529
Ours	2048	5.32	0.529

Table 2: Results on MVP validation set in terms of L2 CD (lower is better) and $F1$ -Score@1% (higher is better) metrics. Inputs and outputs contain 2048 points.

In the comparative results, it is observable that within the category of the lamp, our method generates a point distribution that is more uniform and complete, especially in the details of the lamp, such as the coil part where details are typically missing. The point cloud produced by our method is more precise in these areas. As for the chair category, in the comparison results in Figure 4, we can see that the point distribution on the chair’s back generated by our method is more uniform and complete than other methods.

Results on MVP dataset. Table 2 shows the comparison results of our method with PCN (Yuan et al. 2018), TopNet (Tchapmi et al. 2019), ECG (Pan 2020), CRN (Wang, Ang Jr, and Lee 2020), CRN (Wang, Ang Jr, and Lee 2020), VRCNet (Pan et al. 2021), PoinTr (Yu et al. 2021), CRA-PCN (Rong et al. 2024) on MVP dataset. We can find that our method achieves plausible performance.

Results on KITTI dataset. To demonstrate the effective-

Methods	GRNet	PoinTr	SeedFormer	SVDFormer	Ours
MMD ↓	5.350	32.854	1.179	0.967	0.574

Table 3: Results on KITTI in terms of MMD (lower is better) metric. The baseline is AdaPoinTr (Yu et al. 2023).

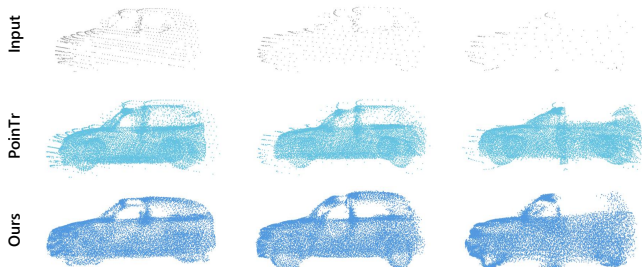


Figure 5: Qualitative results on the KITTI. From the comparison results, our method can obtain more plausible results compared with other work.

ness of our method in real-world scenarios, we evaluate our method on the KITTI dataset (Geiger et al. 2013) and make comparisons with SOTA methods, including, GRNet (Xie et al. 2020), PoinTr (Yu et al. 2021), SeedFormer (Zhou et al. 2022), and SVDFormer (Zhu et al. 2023a). We present the results using the MMD) as a quantitative metric to assess our method without finetuning or retraining like (Zhu et al. 2023a), as detailed in Table 3. The visual comparison in Figure 5 indicates that our method also achieves superior visual results in the task of sparse point cloud completion compared to other methods.

Ablation Study

In this section, we implement ablation studies to demonstrate the effectiveness of the proposed CLIP-enhanced module and position-aware module. All experiments are conducted under unified settings on the PCN dataset.

Generalizability of CLIP-enhanced module. To validate the generalizability, we experiment with two types of text encoders: the text transformer model and the Bert-based model. We find that both models can provide robust results.

Effectiveness of inpainting global projection map. To verify the impact of the inpainting global projection map on point cloud completion, we conduct corresponding ablation studies. We replace the incompleting projection map with the projection map of the complete point cloud. It indicates a significant enhancement in the point cloud completion effect. Specifically, on the CRA-PCN (Rong et al. 2024) baseline, we gain an improvement of nearly 5%, as shown in the last column (CRA-PCN-GT) of Table 4. These comparative results reveal the importance of the complete projection map for point cloud completion. By this contrast experiment, we discover that while the local information provided by the divided images and text description can bring some performance improvement, the absence of global information leads to a certain degree of performance degradation. However, the complete projection map provides more accu-

With / CE			
SnowflakeNet	PoinTr	AdaPoinTr	CRA-PCN
7.19	7.26	6.49	6.37

With / CE + With / PA				
SnowflakeNet	PoinTr	AdaPoinTr	CRA-PCN	CRA-PCN-GT
7.17	7.20	6.48	6.34	6.11

Table 4: Ablation study on PCN dataset in terms of L1 CD $\times 10^3$ (lower is better). CE and PA denote the CLIP-enhanced and Position-aware modules, respectively.

rate positional information, which is beneficial for the model to better predict the missing parts of the point cloud and thus improves the performance of point cloud completion.

Superiority of the proposed framework. To further validate our framework superiority, we conduct four different baselines, SnowflakeNet (Xiang et al. 2021), PoinTr (Yu et al. 2021), AdaPoinTr (Yu et al. 2023) and CRA-PCN (Rong et al. 2024), with the detailed results shown in Table 4. SnowflakeNet is the classic approach that encodes incomplete points into a one-dimensional feature and then decodes the complete point cloud through a coarse-to-fine method. PoinTr and AdaPoinTr redefine the point cloud completion task, considering it a set-to-set translation problem. CRA-PCN represents the current SOTA results. We select these four models as our baselines and successively integrate our proposed CLIP-enhanced and Position-aware modules into these networks. The experimental results indicate that performance improvements of varying degrees are achieved across all baselines. The details are shown in Table 4.

Conclusion

In this paper, we propose a method for efficiently conveying a unimodal point cloud completion framework into a multimodal point cloud completion framework based on the CLIP model. The main motivation of our work is first to introduce a general method that can generate triple multimodal data of point cloud, textual description, and six projection images. Then, each projection map is divided into four equal non-overlapping blocks. We design a position-aware guided module that can learn a local-global weighted map to identify the missing condition in each block. Extensive experiments demonstrate that the proposed dataset and position-aware guided module can improve our method of understanding the semantics and position information of the input point cloud.

In the future, we will investigate the relationship between text descriptions and incomplete point clouds, and develop a fine-grained text-guided framework for 3D point cloud completion. Additionally, we will evaluate our model on large-scale datasets, such as Objaverse, to assess its performance.

Acknowledgments

This work is supported by Beijing Natural Science Foundation (4232023), R&D Program of Beijing Municipal Education Commission (KM202310009002), National Natural Science Foundation of China (62102208), and Yuxiu Innovation Project of NCUT (2024NCUTYXCX202).

References

- Aiello, E.; Valsesia, D.; and Magli, E. 2022. Cross-modal learning for image-guided point cloud shape completion. *NeurIPS*, 37349–37362.
- Cai, P.; Scott, D.; Li, X.; and Wang, S. 2024. Orthogonal Dictionary Guided Shape Completion Network for Point Cloud. In *AAAI*, 864–872.
- Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.
- Chefer, H.; Gur, S.; and Wolf, L. 2021. Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *ICCV*, 397–406.
- Chen, Y.-N.; Dai, H.; and Ding, Y. 2022. Pseudo-stereo for monocular 3d object detection in autonomous driving. In *CVPR*, 887–897.
- Chen, Z.; Long, F.; Qiu, Z.; Yao, T.; Zhou, W.; Luo, J.; and Mei, T. 2023. Anchorformer: Point cloud completion from discriminative nodes. In *CVPR*, 13581–13590.
- Cheng, X.; and Li, L. 2024. Open 3D World in Autonomous Driving. *arXiv preprint arXiv:2408.10880*.
- Christen, S.; Yang, W.; Pérez-D’Arpino, C.; Hilliges, O.; Fox, D.; and Chao, Y.-W. 2023. Learning human-to-robot handovers from point clouds. In *CVPR*, 9654–9664.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houshy, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.
- Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 1231–1237.
- Han, Z.; Chen, C.; Liu, Y.-S.; and Zwicker, M. 2020. ShapeCaptioner: Generative caption network for 3D shapes by learning a mapping from parts detected in multiple views to sentences. In *ACM MM*, 1018–1027.
- Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 4904–4916. PMLR.
- Kasten, Y.; Rahamim, O.; and Chechik, G. 2024. Point cloud completion with pretrained text-to-image diffusion models. *NeurIPS*, 36.
- Li, L. 2023. Hierarchical edge aware learning for 3d point cloud. In *Computer Graphics International Conference*, 81–92. Springer.
- Li, S.; Gao, P.; Tan, X.; and Wei, M. 2023. Proxyformer: Proxy alignment assisted point cloud completion with missing part sensitive transformer. In *CVPR*, 9466–9475.
- Mao, J.; Shi, S.; Wang, X.; and Li, H. 2023. 3D object detection for autonomous driving: A comprehensive survey. *IJCV*, 1909–1963.
- Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; and Ebrahimi, M. 2019. Edgeconnect: Structure guided image inpainting using edge prediction. In *ICCV*.
- Nguyen, D. T.; Hua, B.-S.; Tran, K.; Pham, Q.-H.; and Yeung, S.-K. 2016. A field model for repairing 3d shapes. In *CVPR*, 5676–5684.
- Oehmcke, S.; Li, L.; Revenga, J. C.; Nord-Larsen, T.; Trepekli, K.; Gieseke, F.; and Igel, C. 2022. Deep learning based 3D point cloud regression for estimating forest biomass. In *Proceedings of the 30th international conference on advances in geographic information systems*, 1–4.
- Oehmcke, S.; Li, L.; Trepekli, K.; Revenga, J. C.; Nord-Larsen, T.; Gieseke, F.; and Igel, C. 2024. Deep point cloud regression for above-ground forest biomass estimation from airborne LiDAR. *Remote Sensing of Environment*, 302: 113968.
- Pan, L. 2020. ECG: Edge-aware point cloud completion with graph convolution. *IEEE Robotics and Automation Letters*, 4392–4398.
- Pan, L.; Chen, X.; Cai, Z.; Zhang, J.; Zhao, H.; Yi, S.; and Liu, Z. 2021. Variational relational point completion network. In *CVPR*, 8524–8533.
- Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 652–660.
- Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.
- 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 *AAAI*, 4676–4685.
- Song, W.; Zhou, J.; Wang, M.; Tan, H.; Li, N.; and Liu, X. 2023. Fine-grained Text and Image Guided Point Cloud Completion with CLIP Model. *arXiv preprint arXiv:2308.08754*.
- Tang, J.; Gong, Z.; Yi, R.; Xie, Y.; and Ma, L. 2022. Lake-net: Topology-aware point cloud completion by localizing aligned keypoints. In *CVPR*, 1726–1735.
- Tchapmi, L. P.; Kosaraju, V.; Rezatofighi, H.; Reid, I.; and Savarese, S. 2019. Topnet: Structural point cloud decoder. In *CVPR*, 383–392.
- Wang, J.; Zhou, P.; Shou, M. Z.; and Yan, S. 2023. Position-guided text prompt for vision-language pre-training. In *CVPR*, 23242–23251.
- Wang, X.; Ang Jr, M. H.; and Lee, G. H. 2020. Cascaded refinement network for point cloud completion. In *CVPR*, 790–799.
- Wen, X.; Xiang, P.; Han, Z.; Cao, Y.-P.; Wan, P.; Zheng, W.; and Liu, Y.-S. 2021. Pmp-net: Point cloud completion by learning multi-step point moving paths. In *CVPR*, 7443–7452.
- Wen, X.; Xiang, P.; Han, Z.; Cao, Y.-P.; Wan, P.; Zheng, W.; and Liu, Y.-S. 2022. Pmp-net++: Point cloud completion by transformer-enhanced multi-step point moving paths. *TPAMI*, 852–867.

Xiang, P.; Wen, X.; Liu, Y.-S.; Cao, Y.-P.; Wan, P.; Zheng, W.; and Han, Z. 2021. Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *ICCV*, 5499–5509.

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

Yan, X.; Yan, H.; Wang, J.; Du, H.; Wu, Z.; Xie, D.; Pu, S.; and Lu, L. 2022. Fbnet: Feedback network for point cloud completion. In *ECCV*, 676–693.

Yang, Y.; Feng, C.; Shen, Y.; and Tian, D. 2018. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, 206–215.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Trans. Mach. Learn. Res.*

Yu, X.; Rao, Y.; Wang, Z.; Liu, Z.; Lu, J.; and Zhou, J. 2021. PointTr: Diverse point cloud completion with geometry-aware transformers. In *ICCV*, 12498–12507.

Yu, X.; Rao, Y.; Wang, Z.; Lu, J.; and Zhou, J. 2023. Ada-PoinTr: Diverse Point Cloud Completion With Adaptive Geometry-Aware Transformers. *TPAMI*, 14114–14130.

Yuan, W.; Khot, T.; Held, D.; Mertz, C.; and Hebert, M. 2018. Pcn: Point completion network. In *3DV*, 728–737.

Zhang, B.; Zhao, X.; Wang, H.; and Hu, R. 2022a. Shape completion with points in the shadow. In *SIGGRAPH Asia*, 1–9.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021a. VinVL: Making Visual Representations Matter in Vision-Language Models. *CVPR*.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022b. Pointclip: Point cloud understanding by clip. In *CVPR*, 8552–8562.

Zhang, W.; Yan, Q.; and Xiao, C. 2020. Detail preserved point cloud completion via separated feature aggregation. In *ECCV*, 512–528.

Zhang, W.; Zhou, H.; Dong, Z.; Liu, J.; Yan, Q.; and Xiao, C. 2022c. Point cloud completion via skeleton-detail transformer. *TVCG*, 29(10): 4229–4242.

Zhang, X.; Feng, Y.; Li, S.; Zou, C.; Wan, H.; Zhao, X.; Guo, Y.; and Gao, Y. 2021b. View-guided point cloud completion. In *CVPR*, 15890–15899.

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 *ECCV*, 416–432.

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 *ICCV*, 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. *TVCG*.