Rethinking Multimodal Point Cloud Completion: A Completion-by-Correction Perspective

Wang Luo, Di Wu*, Hengyuan Na, Yinlin Zhu, Miao Hu, Guocong Quan

Sun Yat-sen University, Guangzhou, China
Guangzhou Yunshan Research Institute of Artificial Intelligence Security, Guangzhou, China
luow69@mail2.sysu.edu.cn, wudi27@mail.sysu.edu.cn, neihy@mail2.sysu.edu.cn, zhuylin27@mail2.sysu.edu.cn, humiao5@mail.sysu.edu.cn, quangc@mail.sysu.edu.cn

Abstract

Point cloud completion aims to reconstruct complete 3D shapes from partial observations, which is a challenging problem due to severe occlusions and missing geometry. Despite recent advances in multimodal techniques that leverage complementary RGB images to compensate for missing geometry, most methods still follow a *Completion-by-Inpainting* paradigm, synthesizing missing structures from fused latent features. We empirically show that this paradigm often results in structural inconsistencies and topological artifacts due to limited geometric and semantic constraints. To address this, we rethink the task and propose a more robust paradigm, termed *Completion-by-Correction*, which begins with a topologically complete shape prior generated by a pretrained image-to-3D model and performs feature-space correction to align it with the partial observation. This paradigm shifts completion from unconstrained synthesis to guided refinement, enabling structurally consistent and observation-aligned reconstruction. Building upon this paradigm, we introduce PGNet, a multi-stage framework that conducts dual-feature encoding to ground the generative prior, synthesizes a coarse yet structurally aligned scaffold, and progressively refines geometric details via hierarchical correction. Experiments on the ShapeNetViPC dataset demonstrate the superiority of PGNet over state-of-the-art baselines in terms of average Chamfer Distance (-23.5%) and F-score (+7.1%).

Code — https://github.com/RobWonn/PGNet

1 Introduction

With the popularity of LiDAR and RGB-D cameras, point clouds have emerged as a fundamental 3D representation and are widely adopted in diverse AI applications, including autonomous driving (Chen et al. 2020), augmented reality (Wang et al. 2023), and robotics (Varley et al. 2017). However, point clouds captured by these sensors are often sparse and incomplete due to occlusion, light reflection, and limited resolution, hindering the performance of downstream tasks (Liang et al. 2019; Nie et al. 2021). Consequently, point cloud completion, which aims to recover complete point clouds from partial input, has become an indispensable task.

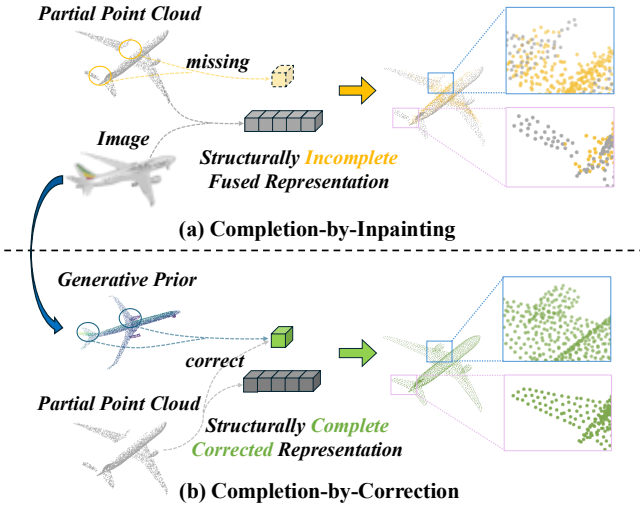

Figure 1: Comparison of two point cloud completion paradigms. (a) Completion-by-Inpainting synthesizes missing geometry from incomplete representation, often introducing artifacts. (b) Our Completion-by-Correction leverages a complete generative prior, correcting it by grounding to the partial observation for more consistent structures.

Traditional methods (Kazhdan and Hoppe 2013; Pauly et al. 2008; Mitra et al. 2013; Berger et al. 2014) for point cloud completion relied on geometric heuristics or template matching, constructing complete shapes by aligning geometric rules or retrieving similar instances from predefined databases. In recent years, deep learning-based unimodal approaches (Yuan et al. 2018; Yu et al. 2021; Cai et al. 2024) have made significant strides by learning shape priors from large-scale datasets and directly synthesizing completions from partial inputs. However, when relying only on geometric input without other contextual cues, it remains difficult to judge whether missing parts stem from occlusion or represent actual voids in the object's structure, which often leads to sub-optimal completion performance.

Motivated by the human capability to perceive 3D structures from 2D views, recent studies explore multimodal learning for point cloud completion, using RGB images as

*Corresponding author.

a readily accessible source of additional texture and semantic information. For example, CSDN (Zhu et al. 2023) treats completion as a style transfer task by injecting image features into a folding-based decoder, and introduces a dual-refinement module leveraging global image guidance and local geometric cues. XMFNet (Aiello, Valsesia, and Magli 2022) uses stacked cross- and self-attention layers to fuse image and point cloud features in the latent space, reconstructing shapes via multiple prediction branches. In contrast, EGIInet (Xu et al. 2024) adopts an explicitly guided interaction strategy with a novel loss to align structural information across modalities before fusion.

Despite their effectiveness, these methods face **Critical Limitations**: As illustrated in Figure 1(a), most existing approaches adopt a *Completion-by-Inpainting* paradigm, synthesizing missing geometry from fused visual and geometric features. We empirically demonstrate that this process is inherently uncertain and becomes unreliable under severe degradation, as the network operates without an explicit structural scaffold and must hallucinate structure from limited guidance (Sec. 4.2). As a result, although the generated completions appear semantically plausible, they frequently exhibit structural inconsistencies and topological artifacts.

To this end, we rethink the task of multimodal point cloud completion and propose a more robust paradigm, termed the *Completion-by-Correction*. As illustrated in Figure 1(b), rather than synthesizing geometry from an incomplete fused representation, we begin with a topologically complete and semantically meaningful shape prior generated by a pretrained image-to-3D model. We subsequently perform feature-space correction to align this prior with the partial observation, instead of relying on direct geometric fusion (Zhang et al. 2021; Wei et al. 2025), which is often undermined by pose and scale misalignment introduced by model bias and image ambiguity. This paradigm shifts the task from unconstrained synthesis to guided refinement over an observation-consistent representation, making it more well-posed and robust to geometric uncertainty.

Building upon this insight, we introduce **P**rior**G**round**Net** (PGNet), a novel framework that realizes the Completion-by-Correction paradigm through a carefully designed three-stage process. (1) *Corrective Dual-Feature Encoding*, the generative prior is corrected by grounding its features in the partial observation via parallel encoding and semantic correspondence; (2) *Grounded Seed Generation*, synthesizes a coarse but topologically complete point cloud that serves as a structural scaffold aligned with the observation; (3) *Hierarchical Grounded Refinement*, iteratively refines the coarse scaffold by aggregating dual-source features for each point, capturing high-fidelity geometry from the observation and structural context from the prior, predicting displacements informed by shape context.

Our Contributions: (1) **Paradigm Identification.** We introduce the Completion-by-Correction paradigm, which reformulates multimodal point cloud completion. Instead of inpainting missing regions, our paradigm corrects a structurally complete generative shape prior by grounding it in the partial observation, reducing geometric ambiguity and structural artifacts. (2) **Novel Framework.** We propose PGNet, a framework that implements our paradigm through three dedicated stages: Corrective Dual-Feature Encoding, Grounded Seed Generation, and Hierarchical Grounded Refinement. (3) **SOTA Performance.** PGNet consistently achieves state-of-the-art performance on the ShapeNetViPC dataset, reducing average Chamfer Distance by 23.5% and improving average F-score by 7.1% over prior methods.

2 Related Works

2.1 Unimodal Point Cloud Completion

Traditional methods include geometry-based and alignment-based techniques. Geometry-based methods (Berger et al. 2014; Davis et al. 2002; Nealen et al. 2006; Sarkar, Varanasi, and Stricker 2017) infer missing structures using geometric priors such as surface smoothness or local symmetry. These approaches typically interpolate or replicate geometric patterns but often struggle with incomplete or complex structures. Alignment-based (Mitra, Guibas, and Pauly 2006; Mitra et al. 2013; Pauly et al. 2008; Sipiran, Gregor, and Schreck 2014; Sung et al. 2015) methods retrieve and fit models from shape databases via shape matching, part assembly, or template deformation. However, their applicability is constrained by computational cost, noise sensitivity, and limited dataset coverage.

Deep learning methods have demonstrated superior performance by learning data-driven shape priors. PCN (Yuan et al. 2018) introduced a PointNet-based encoder and a FoldingNet-style (Yang et al. 2018) decoder, enabling end-to-end completion without strong geometric assumptions. TopNet (Tchapmi et al. 2019) adopted a tree-structured decoder to support hierarchical generation, while PF-Net (Huang et al. 2020) enhanced feature fusion and cascaded refinement. Transformer-based architectures further advanced this line of research: PoinTr (Yu et al. 2021) reformulated completion as set-to-set translation and incorporated a geometry-aware module; SnowFlakeNet (Xiang et al. 2021) applied iterative snowflake-like deconvolution. SeedFormer (Zhou et al. 2022) proposed a shape representation using patch seeds that capture both global structure and local details. PointAttN (Wang et al. 2024) employs attention mechanisms to capture local and global structures among unordered points without explicit region partitioning. CRA-PCN (Rong et al. 2024) and PointCFormer (Zhong et al. 2025) improved local geometric modeling through cross-resolution attention and relation-based metrics. Recent works (Wei et al. 2025; Yan et al. 2025; Kasten, Rahamim, and Chechik 2023; Chu et al. 2025) also explored diffusion models (Ho, Jain, and Abbeel 2020), symmetry priors, and context-aware refinement strategies.

2.2 Multimodal Point Cloud Completion

Multimodal methods leverage auxiliary modalities, typically a single RGB image, to guide point cloud completion. Early works such as ViPC (Zhang et al. 2021) concatenated coarse image-derived point clouds with partial inputs for structural enhancement. More recent methods employ advanced fusion strategies. XMFNet (Aiello, Valsesia, and Magli 2022)

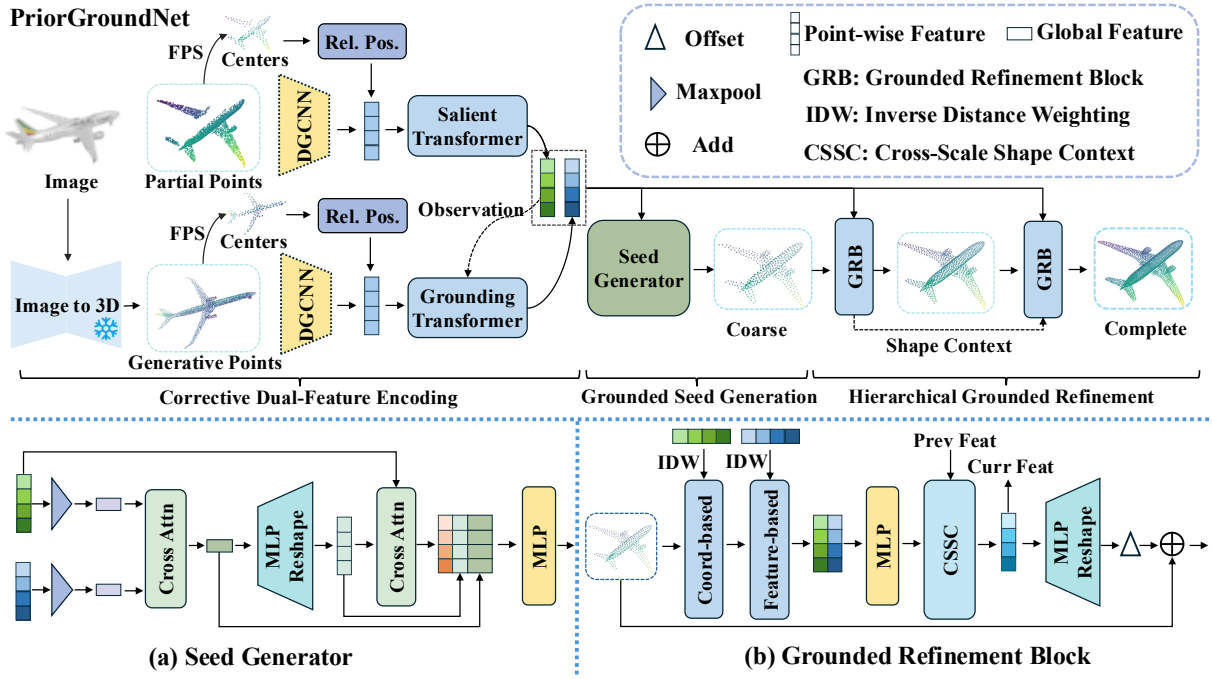

Figure 2: The overall architecture of PriorGroundNet (PGNet), which follows the Completion-by-Correction paradigm and consists of three stages: Corrective Dual-Feature Encoding, Grounded Seed Generation, and Hierarchical Grounded Refinement. (a) The detailed structure of the Seed Generator module, which produces a coarse but complete point cloud by grounding semantic seeds. (b) The architecture of the Grounded Refinement Block (GRB), which hierarchically enhances geometric detail using dual-source feature.

performs implicit fusion using layered attention and reconstructs shapes via multi-branch prediction. CSDN (Zhu et al. 2023) treats completion as a style transfer task by embedding image features into the decoding process and refining output through joint image and geometry cues. EGIInet (Xu et al. 2024) enhances structural understanding through explicitly guided interaction and modality alignment. CDPNet (Du et al. 2024) adopts a patch-based strategy to densify coarse point clouds and extracts a cross-modal style code to guide structural detail generation. DMF-Net (Mao et al. 2025) introduces a dual-channel fusion framework with a shape-aware upsampling transformer, enabling balanced exploitation of image and point cloud features.

In contrast to prior works that synthesize missing geometry directly from fused features, we reformulate completion as a correction process over a complete shape prior. By grounding this prior in partial observations, our approach transforms the task from uncertain generation to observation-guided structural refinement.

3 Methodology

3.1 Overview

The overall architecture of PGNet is shown in Figure 2, which consists of three stages: Corrective Dual-Feature Encoding, Grounded Seed Generation, and Hierarchical Grounded Refinement. We will detail each module in the following.

Problem Definition. Given a partial point cloud observation $P_o \in \mathbb{R}^{M\times 3}$ and its corresponding single-view RGB image $I \in \mathbb{R}^{H\times W\times 3}$, the objective of multimodal point cloud completion is to reconstruct a complete and high-fidelity point cloud $P_{gt} \in \mathbb{R}^{N\times 3}$. Here, M and N denote the number of points in the partial and ground truth point clouds, respectively.

The Completion-by-Correction Paradigm. More robust and well-posed, the Completion-by-Correction paradigm replaces ill-posed synthesis with the correction of a generative shape prior, guided by partial observations. It initializes from a topologically complete and semantically meaningful shape prior $P_g = \mathcal{G}(I)$, generated by a pre-trained image-to-3D model $\mathcal{G}$. While $P_g \in \mathbb{R}^{N_g\times 3}$ provides a strong structural scaffold, it may contain geometric inaccuracies due to model bias or input ambiguity. The goal is thus reframed as learning a correction function $\mathcal{T}$ that grounds P_g in the partial observation P_o, yielding a final output $P = \mathcal{T}(P_g, P_o)$ that aligns with the ground truth P_{gt}.

3.2 Corrective Dual-Feature Encoding

A key challenge in this paradigm is the discrepancy between the generative prior P_g and the partial observation P_o, which often differ in scale, pose, and point distribution and may contain hallucinated geometry. We therefore encode P_g and P_o in parallel and ground the prior features on those from the partial observation in feature space, enabling robust align-

ment and providing a corrected context for subsequent generation.

Partial Point Cloud Encoder. The partial point cloud P_o provides reliable geometric evidence. As illustrated in Figure 2(a), we adopt a hierarchical local feature aggregation strategy following PoinTr (Yu et al. 2021): Farthest Point Sampling (FPS) selects N_e representative points, and a lightweight DGCNN (Wang et al. 2019) aggregates local features around each point. This yields point centers $C_o = \{c_{o,i}\}_{i=1}^{N_e}$ and an initial feature tensor $F'_o \in \mathbb{R}^{N_e \times D}$.

To alleviate pose and scale discrepancies between P_o and P_g, we use a learnable relative position embedding Φ that encodes the local spatial layout of each patch and is added to the initial features:

$$F''_o = F'_o + \Phi(C_o). \tag{1}$$

While DGCNN effectively captures local geometric structure, it is limited in modeling long-range dependencies and handling non-uniform point densities. We therefore introduce the Salient Transformer, a dual-branch architecture that integrates global and local context. The global branch applies multi-head self-attention (MHSA) to F''_o to produce long-range context features A_o, and the local branch aggregates k-nearest-neighbor patterns X_o via shared MLPs and max pooling. To adaptively combine the two sources of information, a learnable salience gate dynamically fuses global and local features:

$$\begin{aligned} G_o &= \sigma(\text{MLP}([A_o, X_o])), \\ F_o &= (1 - G_o) \odot A_o + G_o \odot X_o, \end{aligned} \tag{2}$$

where σ denotes the sigmoid function and $\odot$ denotes element-wise multiplication. This mechanism allows the network to emphasize global context in sparse regions and local detail where fine geometry is critical.

Generative Point Cloud Encoder. We apply the same hierarchical encoding to P_g, obtaining representative centers C_g and initial features $F'_g \in \mathbb{R}^{N_e \times D}$. Adding the shared positional encoding yields

$$F''_g = F'_g + \Phi(C_g). \tag{3}$$

On top of F''_g, we build a Grounding Transformer that refines the generative prior in feature space under the guidance of the partial observation. It processes F''_g through two parallel branches: a self-attention branch that applies MHSA to produce contextual features A_g encoding internal structural priors of P_g, and a grounding branch that performs cross-attention with F''_g as queries and F_o as keys and values to obtain observation-aligned features X_g. To adaptively balance these complementary cues, we reuse the same salience gate formulation as in the Salient Transformer:

$$\begin{aligned} G_g &= \sigma(\text{MLP}([A_g, X_g])), \\ F_g &= (1 - G_g) \odot A_g + G_g \odot X_g, \end{aligned} \tag{4}$$

The resulting features $F_g = \{f_{g,i}\}_{i=1}^{N_e}$ form a corrected, observation-aware representation of the generative prior, which serves as the foundation for subsequent grounded seed generation and hierarchical refinement. Conceptually, the Salient Transformer enhances the reliability of F_o by balancing global and local evidence, while the Grounding Transformer injects this reliable observation signal into the generative prior.

3.3 Grounded Seed Generation

The goal of this stage is to synthesize a structurally complete yet geometrically grounded scaffold as the basis for subsequent refinement. Consistent with our Completion-by-Correction paradigm, we leverage the topological completeness of the generative prior and align it with the geometric fidelity of the partial observation.

As illustrated in Figure 2(a), we first perform max pooling on F_g and F_o to extract their global representations $\hat{F}_g$ and $\hat{F}_o$. A cross-attention module is then applied, with $\hat{F}_o$ as the query and $\hat{F}_o$ as key and value, yielding a fused global feature $\hat{F}_{\text{fused}}$ that encodes an observation-aware yet topologically complete shape representation.

Next, instead of directly regressing point coordinates from $\hat{F}_{\text{fused}}$, we generate a structured set of N_c seed features $F_{\text{seed}} \in \mathbb{R}^{N_c \times D}$ by projecting $\hat{F}_{\text{fused}}$ through an MLP and reshaping the output. This operation, inspired by the PixelShuffle (Shi et al. 2016) mechanism, effectively expands the global feature into a spatially organized seed representation, enabling the network to model inter-point dependencies and produce coherent, structurally consistent layouts.

To further incorporate geometric grounding, we apply cross-attention with F_{seed} as the query and F_o as the key and value, producing grounded features F_{gr}. The final coarse point cloud P_c is then obtained by feeding the concatenation of the replicated global feature, seed features, and grounded features into an MLP:

$$P_c = \text{MLP}([\text{Replicate}(\hat{F}_{\text{fused}}, N_c), F_{\text{seed}}, F_{\text{gr}}]). \tag{5}$$

In this way, global priors propose a complete seed layout, while point-wise grounding from F_o adjusts it towards the observed geometry, yielding a structurally complete and observation-aware scaffold for later refinement.

3.4 Hierarchical Grounded Refinement

Unlike inpainting-based methods that decode abstract features into dense point sets, our approach refines geometry using corrected features from feature-space grounding. The coarse output P_c offers a complete but low-resolution scaffold aligned with the observation. Built on this, we introduce a hierarchical architecture of K stacked Grounded Refinement Blocks (GRBs) (Figure 2(b)), which progressively improve geometric fidelity via localized matching guided by both the observation and corrected prior.

Each GRB contains two components: (1) Dual-Source Feature Association, which retrieves semantically aligned local patterns from both sources; and (2) Structure-Aware Upsampling, which captures shape context and predicts localized displacements to refine spatial layout.

Dual-Source Feature Association. Given an input point set $P_{\text{in}} \in \mathbb{R}^{N_{\text{in}} \times 3}$ from the previous stage, we enrich each point p_i with localized features by querying both the partial observation and the corrected generative prior. This association propagates observation guidance while injecting structurally plausible patterns from the prior.

We begin by querying features from the partial observation. Given the center points C_o and their associated features F_o, we employ inverse distance weighting (IDW) (Qi et al. 2017) to interpolate features for each query point p_i based on its k nearest neighbors in C_o:

$$\begin{aligned} f_{\text{interp},o}(p_i) &= \frac{\sum_{j \in \mathcal{N}_k(p_i, C_o)} w_j f_{o,j}}{\sum_{j \in \mathcal{N}_k(p_i, C_o)} w_j}, \\ w_j &= \frac{1}{\|p_i - c_{o,j}\|_2}, \end{aligned} \tag{6}$$

where $\mathcal{N}_k(p_i, C_o)$ denotes the k nearest neighbors of p_i in Euclidean space. To integrate prior structure, we additionally query from F_g. As direct spatial interpolation may fail due to misalignment between P_o and P_g, we instead interpolate in the feature space. Specifically, each $f_{\text{interp},o}(p_i)$ is used to find its k nearest neighbors in F_g, and IDW is applied based on feature distance:

$$\begin{aligned} f_{interp,g}(p_i) &= \frac{\sum_{j \in \mathcal{N}_k(f_{interp,o}(p_i), F_g)} w'_j f_{g,j}}{\sum_{j \in \mathcal{N}_k(f_{interp,o}(p_i), F_g)} w'_j}, \\ w'_j &= \frac{1}{\|f_{interp,o}(p_i) - f_{g,j}\|_2}. \end{aligned} \tag{7}$$

The final dual-source representation is formed by concatenating both:

$$f_{as}(p_i) = [f_{interp,o}(p_i), f_{interp,g}(p_i)]. \tag{8}$$

This dual-source association is critical for resolving coordinate misalignment and grounding geometric features in partial observations. Structural patterns from the generative prior are selectively integrated under observation guidance, ensuring that only semantically consistent regions contribute to refinement.

Structure-Aware Upsampling. Given the fused dual-source features for each point in P_{in}, we aim to refine geometry by enhancing structural fidelity and increasing point density. To this end, we aggregate multi-scale shape context so that each point receives geometric information from its local neighborhood in the previous resolution, enabling more informed prediction of point-wise displacements.

Central to this stage is the Cross-Scale Shape Context (CSSC) module, which enhances each point in the current resolution by attending to geometrically relevant features from a source point set. This source consists of spatial coordinates P_k and their associated features, derived from the contextual outputs F_{ctx}^{prev} of the previous refinement stage. Note that in the first block, where no prior context is available, both P_k and the associated features default to the input point set P_{in} and its projected features. To support cross-scale attention, the fused dual-source features F_{as} are projected to obtain a set of query features F_q. For each query point $p_i \in P_{\text{in}}$, we identify its k nearest neighbors in P_k, and project the corresponding features into key and value vectors k_j and v_j, while $f_{q,i}$ is projected into a query vector q_i. Following geometric transformer design (Zhao et al. 2021), attention weights are computed based on both feature similarity and relative spatial position:

$$\alpha_{ij} = \frac{\exp(\text{MLP}(q_i - k_j + \Phi(p_i - p_{k,j})))}{\sum_{l \in \mathcal{N}_k(p_i, P_k)} \exp(\text{MLP}(q_i - k_l + \Phi(p_i - p_{k,l})))} \tag{9}$$

where Φ denotes the relative positional encoding. The final contextual feature $f_{ctx,i}$ is computed by weighted aggregation of neighbor values, with a residual connection:

$$f_{ctx,i} = f_{q,i} + \sum_{j \in \mathcal{N}_k(p_i)} \alpha_{ij} v_j. \tag{10}$$

This CSSC module enriches each point with geometric structure from the previous resolution, facilitating spatially consistent upsampling. Once the contextual features $F_{ctx} = \{f_{ctx,i}\}_{i=1}^{N_{\text{in}}}$ are computed for all input points, they are used to predict displacements for upsampling. An MLP first expands the channel dimension of F_{ctx} and the result is reshaped to generate r distinct displacement vectors for each point, which can be formulated as

$$\Delta = \text{Reshape}(\text{MLP}(F_{ctx})), \tag{11}$$

where $\Delta \in \mathbb{R}^{rN_{\text{in}} \times 3}$ represents the complete set of predicted offsets and r is the upsampling rate. The final upsampled point cloud P_{up} is obtained by replicating the input point set and adding the predicted displacements:

$$P_{up} = \text{Replicate}(P_{\text{in}}, r) + \Delta. \tag{12}$$

3.5 Loss Function

Following prior work (Aiello, Valsesia, and Magli 2022; Mao et al. 2025), we adopt the L1 Chamfer Distance (L1-CD) as the training objective. Given a predicted point cloud P_{pred} and the ground truth P_{gt}, the L1-CD is defined as:

$$\begin{aligned} \mathcal{L}_{\text{CD}}(P_{pred}, P_{gt}) = &\frac{1}{|P_{pred}|} \sum_{x \in P_{pred}} \min_{y \in P_{gt}} |x - y|_1 \\ &+ \frac{1}{|P_{gt}|} \sum_{y \in P_{gt}} \min_{x \in P_{pred}} |y - x|_1, \end{aligned} \tag{13}$$

where $|\cdot|_1$ denotes the L1 norm. The first term encourages predicted points to align closely with the ground truth, while the second term ensures complete coverage of the ground truth by the prediction.

To promote consistency across refinement stages, we directly supervise the coarse output P_c and each upsampled point cloud $\{P^{(k)}\}_{k=1}^{K}$ using the high-resolution ground truth P_{gt}. This encourages intermediate predictions to approximate the target distribution early on, facilitating convergence and improving geometric fidelity throughout the hierarchy. The final loss is computed as the unweighted average of the L1-CD values across all prediction levels:

$$\mathcal{L} = \frac{1}{K+1} \left(\mathcal{L}_{\text{CD}}(P_c, P_{gt}) + \sum_{k=1}^{K} \mathcal{L}_{\text{CD}}(P^{(k)}, P_{gt}) \right). \tag{14}$$

4 Experiments

In this section, we present a comprehensive evaluation of PGNet. We first describe the dataset and implementation details (Sec.4.1). Then, we aim to address the following questions: **Q1**: How does PGNet compare with state-of-the-art methods (Sec.4.2)? **Q2**: How does the proposed Completion-by-Correction paradigm compare with Completion-by-Inpainting strategies (Sec.4.3)? **Q3**: How important are individual components of PGNet to its overall performance (Sec.4.4)? Moreover, we provide further experimental investigations about the generalization performance (**Q4**) and the robustness to variations in the generative prior (**Q5**) of PGNet in the supplementary material.

4.1 Dataset and Implementation Details

Dataset. Following prior works (Aiello, Valsesia, and Magli 2022; Xu et al. 2024), we train and evaluate on the ShapeNet-ViPC (Zhang et al. 2021) dataset, comprising 38,328 objects across 13 categories. Each sample consists of a partial input ($M = 2048$ points), its corresponding image, and a ground truth completion ($N = 2048$ points).

Implementation Details. The network is trained end-to-end in PyTorch (Paszke et al. 2019) using the AdamW optimizer with an initial learning rate of 2×10^{-4} and a cosine annealing schedule. Each of the eight categories is trained separately for 100,000 iterations with a batch size of 192 on NVIDIA RTX 4090 GPU. To generate priors, we use the Trellis (Xiang et al. 2025) image-to-3D model to predict a mesh from each input image and apply Poisson disk sampling on the mesh surface to extract 2048 points.

The network architecture adopts DGCNN-based encoders that extract $N_e = 128$ local feature centers. The Seed Generator produces a 512-point coarse scaffold. The encoding modules apply 6-head attention with 768-dimensional hidden layers for semantic representation learning, while other attention modules use 4 heads and 256 dimensions. All non-attention layers use $D = 256$ feature dimensions, and $k = 8$ for feature interpolation during refinement. The hierarchical refinement stage includes $K = 2$ Grounded Refinement Blocks (GRBs), each with an upsampling factor of $r = 2$, progressively generating 1024 and finally 2048 points.

4.2 Performance Comparison (*Answer for* Q1)

To answer **Q1**, we compare PGNet with state-of-the-art methods on the ShapeNet-ViPC dataset, using Chamfer Distance (CD) and F-score as evaluation metrics. As shown in Table 2 and Table 3, our method establishes a new state-of-the-art across all categories, significantly outperforming the previous best method EGIInet with a 23.5% reduction in average CD and a 7.1% improvement in F-score. The gains are particularly pronounced in categories plagued by severe self-occlusion and challenging geometries, such as cabinet and sofa, as our paradigm robustly reconstructs large missing structures by correcting a complete prior rather than hallucinating them from sparse cues. This advantage is clearly illustrated in Figure 3, where PGNet exhibit superior structural integrity and point uniformity.

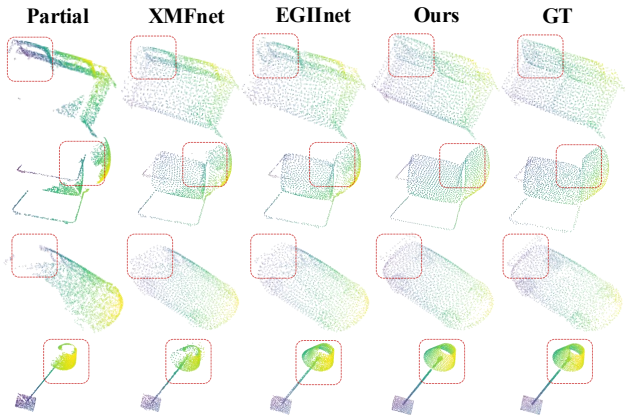

Figure 3: Qualitative comparison on ShapeNet-ViPC.

Method	Avg	Air	Cab	Car	Cha	Lam	Sof	Tab	Wat
Inpaint	1.10	0.53	1.57	1.48	1.21	0.65	1.36	1.29	0.71
PGNet	**0.93**	**0.46**	**1.11**	**1.30**	**1.04**	**0.58**	**1.14**	**1.17**	**0.62**

Table 1: Paradigm-level comparison on ShapeNet-ViPC using CD $\times 10^{-3}$ (lower is better). "Inpaint" represents our inpainting variant.

Unlike other approaches that often produce artifacts or inconsistent geometry, our method generates clean and coherent structures. Specifically, PGNet robustly recovers entire surfaces (e.g., chair seat) and delicate details (e.g., sofa pillows and armrests) by correcting a complete prior scaffold and leveraging dual-source features for refinement.

4.3 Paradigm Comparison (*Answer for* Q2)

To answer **Q2**, we validate our Completion-by-Correction paradigm against a strong inpainting baseline. This baseline replaces our generative prior branch with a pretrained ResNet-18 encoder, while all other components and trainingsettings remain identical. As shown in Table 1, reverting to inpainting significantly degrades performance, increasing average CD by 18.3% (from 0.93 to 1.10), with severe drops in occluded categories such as cabinet (+41.4%). This underscores the limitations of synthesizing geometry directly from fused features and affirms our core premise: shifting the task from unconstrained synthesis to guided correction of a complete scaffold is a more robust strategy.

4.4 Ablation Study (*Answer for* Q3)

To answer **Q3**, we conduct an ablation study on the cabinet category to assess the impact of key PGNet components (Table 4). Removing either feature-level correction (w/o Prior Feature Grounding) or scaffold alignment (w/o Seed Grounding) severely degrades performance (CD: 1.185 and 1.219), highlighting the need for an observation-consistent representation prior to refinement. Disabling dual-source association (w/o Dual-Source Association) causes the largest drop (CD: 1.324), underscoring its role in integrating high-fidelity observation geometry and structural context from

Methods	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
Unimodal Methods									
FoldingNet (Yang et al. 2018)	6.271	5.242	6.958	5.307	8.823	6.504	6.368	7.080	3.882
PCN (Yuan et al. 2018)	5.619	4.246	6.409	4.840	7.441	6.331	5.668	6.508	3.510
TopNet (Tchapmi et al. 2019)	4.976	3.710	5.629	4.530	6.391	5.547	5.281	5.381	3.350
PoinTr (Yu et al. 2021)	2.851	1.686	4.001	3.203	3.111	2.928	3.507	2.845	1.737
SeedFormer (Zhou et al. 2022)	2.902	1.716	4.049	3.392	3.151	3.226	3.603	2.803	1.679
PointAttN (Wang et al. 2024)	2.853	1.613	3.969	3.257	3.157	3.058	3.406	2.787	1.872
Multimodal Methods									
ViPC (Zhang et al. 2021)	3.308	1.760	4.558	3.138	2.476	2.867	4.481	4.990	2.197
CSDN (Zhu et al. 2023)	2.570	1.251	3.670	2.977	2.835	2.554	3.240	2.575	1.742
XMFnet (Aiello, Valsesia, and Magli 2022)	1.454	0.628	1.938	1.753	1.404	1.818	1.748	1.449	0.894
EGIInet (Xu et al. 2024)	1.211	0.552	1.922	1.659	1.203	0.777	1.552	1.227	0.803
Ours	**0.926**	**0.455**	**1.111**	**1.303**	**1.038**	**0.578**	**1.139**	**1.167**	**0.615**

Table 2: Quantitative comparison on ShapeNet-ViPC dataset using Chamfer Distance $\times 10^{-3}$ (lower is better).

Methods	Avg	Airplane	Cabinet	Car	Chair	Lamp	Sofa	Table	Watercraft
Unimodal Methods									
FoldingNet (Yang et al. 2018)	0.331	0.432	0.237	0.300	0.204	0.360	0.249	0.351	0.518
PCN (Yuan et al. 2018)	0.407	0.578	0.270	0.331	0.323	0.456	0.293	0.431	0.577
TopNet (Tchapmi et al. 2019)	0.467	0.593	0.358	0.405	0.388	0.491	0.361	0.528	0.615
PoinTr (Yu et al. 2021)	0.683	0.842	0.516	0.545	0.662	0.742	0.547	0.723	0.780
SeedFormer (Zhou et al. 2022)	0.688	0.835	0.551	0.544	0.668	0.777	0.555	0.716	0.786
PointAttN (Wang et al. 2024)	0.662	0.841	0.483	0.515	0.638	0.729	0.512	0.699	0.774
Multimodal Methods									
ViPC (Zhang et al. 2021)	0.591	0.803	0.451	0.512	0.529	0.706	0.434	0.594	0.730
CSDN (Zhu et al. 2023)	0.695	0.862	0.548	0.560	0.669	0.761	0.557	0.729	0.782
XMFnet (Aiello, Valsesia, and Magli 2022)	0.797	0.957	0.671	0.696	0.809	0.791	0.719	0.823	0.910
EGIInet (Xu et al. 2024)	0.836	0.969	0.691	0.723	0.847	0.919	0.756	0.857	0.927
Ours	**0.895**	**0.985**	**0.839**	**0.804**	**0.887**	**0.954**	**0.850**	**0.881**	**0.963**

Table 3: Quantitative comparison on ShapeNet-ViPC dataset using F-score@0.001 (higher is better).

Model Variant	**CD** $\downarrow$	**F-score** $\uparrow$
w/o Prior Feature Grounding	1.185	0.827
w/o Seed Grounding	1.219	0.821
w/o Dual-Source Association	1.324	0.803
w/o Structure-Aware	1.275	0.800
PGNet (Full Model)	**1.111**	**0.839**

Table 4: Ablation study of PGNet on the ShapeNet-ViPC cabinet category.

the prior. Excluding the structure-aware upsampling module (w/o Structure-Aware) also harms detail recovery (CD: 1.275), confirming the value of shape-guided displacement prediction for local fidelity. In summary, these modules and their components play a pivotal role.

5 Conclusion

In this work, we revisited the Completion-by-Inpainting paradigm for multimodal point cloud completion, which often suffers from structural artifacts when synthesizing missing geometry from limited features. We proposed Completion-by-Correction, a more robust alternative that reframes completion as the guided refinement of a complete generative prior. Instead of direct synthesis from an incomplete fused representation, our method grounds the prior in partial observations through feature-space correction, reducing ambiguity and improving consistency. We introduced PGNet, a three-stage framework that implements this paradigm via corrective encoding, seed generation, and hierarchical refinement. Experiments on ShapeNet-ViPC show that PGNet achieves state-of-the-art accuracy and structural quality. In the future, we will to extend this approach to real-world scenarios using large-scale image-to-3D models with broader category and scene coverage.

Acknowledgments

This work was supported in part by the Shenzhen Science and Technology Program under Grant KJZD20230923113901004, in part by the National Natural Science Foundation of China under Grants 62572501 and 62502551, and in part by the Guangzhou Yunshan Research Institute of Artificial Intelligence Security under Grant HT-99982025-0734. (Corresponding author: Di Wu.)

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.

Berger, M.; Tagliasacchi, A.; Seversky, L.; Alliez, P.; Levine, J.; Sharf, A.; and Silva, C. 2014. State of the art in surface reconstruction from point clouds. *Eurographics 2014-State of the Art Reports*, 1(1): 161–185.

Cai, P.; Scott, D.; Li, X.; and Wang, S. 2024. Orthogonal dictionary guided shape completion network for point cloud. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 864–872.

Chen, S.; Liu, B.; Feng, C.; Vallespi-Gonzalez, C.; and Wellington, C. 2020. 3d point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception. *IEEE Signal Processing Magazine*, 38(1): 68–86.

Chu, J.; Li, W.; Wang, X.; Ning, K.; Lu, Y.; and Fan, X. 2025. Digging into Intrinsic Contextual Information for High-fidelity 3D Point Cloud Completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2573–2581.

Davis, J.; Marschner, S. R.; Garr, M.; and Levoy, M. 2002. Filling holes in complex surfaces using volumetric diffusion. In *Proceedings. First international symposium on 3d data processing visualization and transmission*, 428–441. IEEE.

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

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

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.

Kasten, Y.; Rahamim, O.; and Chechik, G. 2023. Point cloud completion with pretrained text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36: 12171–12191.

Kazhdan, M.; and Hoppe, H. 2013. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3): 1–13.

Liang, M.; Yang, B.; Chen, Y.; Hu, R.; and Urtasun, R. 2019. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7345–7353.

Mao, A.; Tang, Y.; Huang, J.; and He, Y. 2025. DMF-Net: Image-guided point cloud completion with dual-channel modality fusion and shape-aware upsampling transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6063–6071.

Mitra, N. J.; Guibas, L. J.; and Pauly, M. 2006. Partial and approximate symmetry detection for 3d geometry. *ACM Transactions on Graphics (ToG)*, 25(3): 560–568.

Mitra, N. J.; Pauly, M.; Wand, M.; and Ceylan, D. 2013. Symmetry in 3d geometry: Extraction and applications. In *Computer graphics forum*, volume 32, 1–23. Wiley Online Library.

Nealen, A.; Igarashi, T.; Sorkine, O.; and Alexa, M. 2006. Laplacian mesh optimization. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, 381–389.

Nie, Y.; Hou, J.; Han, X.; and Nießner, M. 2021. Rfd-net: Point scene understanding by semantic instance reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4608–4618.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pauly, M.; Mitra, N. J.; Wallner, J.; Pottmann, H.; and Guibas, L. J. 2008. Discovering structural regularity in 3D geometry. In *ACM SIGGRAPH 2008 papers*, 1–11.

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.

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*, volume 38, 4676–4685.

Sarkar, K.; Varanasi, K.; and Stricker, D. 2017. Learning quadrangulated patches for 3d shape parameterization and completion. In *2017 International Conference on 3D Vision (3DV)*, 383–392. IEEE.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Sipiran, I.; Gregor, R.; and Schreck, T. 2014. Approximate symmetry detection in partial 3d meshes. In *Computer graphics forum*, volume 33, 131–140. Wiley Online Library.

Sung, M.; Kim, V. G.; Angst, R.; and Guibas, L. 2015. Data-driven structural priors for shape completion. *ACM Transactions on Graphics (TOG)*, 34(6): 1–11.

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. IEEE.

Wang, J.; Cui, Y.; Guo, D.; Li, J.; Liu, Q.; and Shen, C. 2024. Pointattn: You only need attention for point cloud completion. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 38, 5472–5480.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

Wang, Z.; Nguyen, C.; Asente, P.; and Dorsey, J. 2023. PointShopAR: Supporting environmental design prototyping using point cloud in augmented reality. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–15.

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.

Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2025. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21469–21480.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, 5499–5509.

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.

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*, volume 39, 9094–9102.

Yang, Y.; Feng, C.; Shen, Y.; and Tian, D. 2018. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 206–215.

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.

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.

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.

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*, volume 39, 10689–10697.

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