SRCNet: Robust Point Cloud Completion via Semantic-Guided Part Augmentation

Fei Hu
College of Computer Science
Sichuan University
Chengdu, China
2024223045111@stu.scu.edu.cn

Li Lu*
College of Computer Science
Sichuan University
Chengdu, China
luli@scu.edu.cn

***Abstract*—Point cloud completion is a fundamental challenge in 3D vision, aiming to recover complete geometries from sparse, incomplete sensor data. However, existing methods, predominantly trained on synthetic datasets with uniform viewpoint-based occlusions, exhibit a critical structural blindness. They tend to overfit to specific projection patterns and fail to generalize to the diverse structural defects common in real-world scenarios. To address this, we propose a synergistic framework that enforces structural invariance through both data and architecture. First, we introduce PartDrop, a novel semantic-guided augmentation strategy. Unlike generic erasing techniques that operate on random spatial regions, PartDrop leverages semantic priors to stochastically remove logical components, compelling the network to learn robust part-to-whole dependencies rather than shallow template matching. Second, to effectively reason about these disjointed parts, we propose the Structural-Relational Completion Network (SRCNet). Designed as a specialized Transformer architecture, SRCNet features a structure-aware dual-stream encoder to disentangle semantic context from local topology, and a Region-Relation Decoder equipped with an Inter-region Relation Learning mechanism. By leveraging the inherent global attention capabilities of Transformers, this architecture explicitly models long-range dependencies between observed and missing regions, enabling high-fidelity reconstruction even under severe structural incompleteness. Extensive experiments on PCN, ShapeNet-55, and real-world KITTI benchmarks demonstrate that our method achieves state-of-the-art performance, showing significantly improved robustness and consistency across diverse incompleteness patterns compared to existing approaches.**

***Index Terms*—Point Cloud Completion, Data Augmentation, Transformer, Robustness**

I. Introduction

Point cloud completion is a fundamental challenge in 3D vision, essential for bridging the gap between sparse sensor data and downstream applications. While recent Transformer-based approaches [1], [2] have achieved impressive fidelity on standard benchmarks, they exhibit a critical vulnerability: **a lack of structural robustness**. Existing models are predominantly trained on synthetic datasets with uniform, viewpoint-based occlusions [3]. Consequently, they tend to overfit to specific projection patterns rather than learning the underlying topological logic of objects. As shown in Figure 1, models capable of filling a simple hole often fail catastrophically when a semantic part (e.g., a wing) is missing, revealing that they rely on shallow surface continuity rather than holistic structural reasoning.

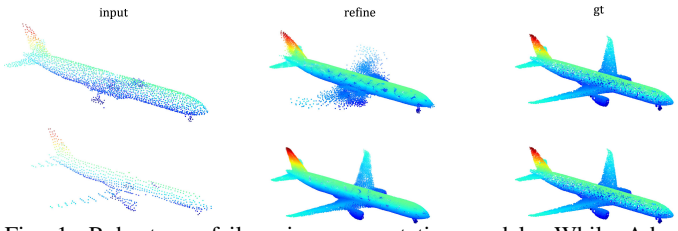

Fig. 1: Robustness failure in representative models. While AdaPoinTr [4] handles standard viewpoint occlusions (Bottom), it collapses when facing structural incompleteness (Top, missing wing), a scenario our method is designed to solve.

To address this, we propose a synergistic framework that enforces **Structural Invariance** through both data and architecture.

First, we introduce **PartDrop**, a novel semantic-guided augmentation strategy. Unlike generic augmentation techniques like Cutout or Random Erasing which operate on spatial regions, PartDrop leverages semantic segmentation priors to stochastically remove logical components. This counterfactual training paradigm forces the model to abandon template matching and instead learn robust part-to-whole dependencies, significantly narrowing the domain gap between synthetic training and diverse real-world structural defects.

Second, to effectively reason about these disjointed parts, we propose a specialized network architecture, the **Structural-Relational Completion Network** (referred to as **SRCNet**). Unlike standard Transformers that process points as a uniform sequence, SRCNet is explicitly designed for structural inference. It features a **Structure-Aware Dual-Stream Encoder** that disentangles semantic context from local topology, and a **Region-Relation Decoder** utilizing an Inter-region Relation Learning (IRL) mechanism. By interacting via representative regional proxies, our model efficiently captures long-range dependencies between observed and missing parts, enabling high-fidelity reconstruction even under severe structural incompleteness.

Our main contributions are:

- We identify the "structural blindness" in current representative models and propose **PartDrop**, a semantic augmentation strategy that significantly boosts robustness against non-uniform defects.

- We present **SRCNet**, a novel architecture integrating a Dual-Stream Encoder and an IRL Decoder to master complex structural reasoning.
- Extensive experiments on PCN, ShapeNet-55, and real-world KITTI benchmarks demonstrate that our method achieves state-of-the-art performance, particularly in recovering fine-grained details and generalizing to unseen structural missing patterns.

II. Related Work

A. Traditional Approaches

Early research on shape completion was primarily driven by geometric rules or database retrieval. Geometry-based methods [5], [6] typically employed surface smoothing algorithms (e.g., Poisson reconstruction) or symmetry detection [7] to infer missing regions. However, these approaches rely heavily on the assumption that the partial input contains sufficient geometric cues, failing when faced with large-scale incompleteness. Alignment-based methods [8], [9] addressed this by retrieving complete templates from shape databases and deforming them to match the input. While effective for common objects, their performance is strictly bounded by the diversity of the database and the accuracy of retrieval, limiting generalization to novel categories.

B. Deep Learning-Based Methods

The introduction of PointNet [10] revolutionized 3D processing by enabling direct operation on unstructured point sets.

a) Point-Based Architectures: To overcome the resolution limits of the voxel-based methods [11], point-based networks adopted encoder-decoder frameworks. PCN [3] pioneered a coarse-to-fine generation strategy, while FoldingNet [12] introduced a 2D-to-3D folding operation to preserve surface continuity. Subsequent work incorporated Generative Adversarial Networks (GANs) [13] to enhance realism or Graph Convolutional Networks (GCNs) [14] to better capture local topology. Despite these advances, convolution-based methods often struggle to model long-range dependencies between spatially distant parts.

b) Transformer-Based Architectures: Recently, Transformers have become the state-of-the-art due to their superior global context modeling. PoinTr [1] reformulated completion as a set-to-set translation problem, utilizing geometry-aware proxies to predict missing points. SeedFormer [2] introduced the concept of patch seeds to generate fine-grained details. More recently, PointAttn [15] optimized the attention mechanism for efficiency, and PCDreamer [16] explored multi-view diffusion priors for high-fidelity generation.

However, a critical limitation persists: Most existing Transformer models treat the point cloud as a uniform sequence of points or patches, applying attention mechanisms that are agnostic to the object's semantic structure. This *point-level* or *patch-level* interaction often fails to capture explicit **part-to-whole** logic (e.g., inferring a wing from a fuselage). Furthermore, these models are predominantly trained on synthetic datasets with fixed-viewpoint occlusions, leading to poor robustness against the diverse structural defects found in real-world scans. In contrast, our **SRCNet** introduces *region-level* relational reasoning to explicitly model structural topology, and our **PartDrop** strategy enforces robustness against structural incompleteness, effectively bridging this gap.

III. Methodology

To address the generalization failure caused by uniform viewpoint training, we propose a robust completion framework that enforces structural invariance through both data augmentation and architectural design. As illustrated in Figure 2, our pipeline consists of two synergistic components: **PartDrop**, a semantically guided augmentation strategy that synthesizes diverse structural defects to mitigate data bias; and **SRCNet** (Structural-Relational Network), a specialized architecture designed to reason about part-to-whole relationships from these complex inputs.

A. PartDrop Augmentation

Existing methods typically rely on viewpoint-based projection to generate partial data, which limits models to learning surface continuity rather than structural logic. To overcome this, as visualized in the input stage of Figure 2, we introduce **PartDrop**, a semantic-guided augmentation strategy. Unlike generic techniques like Cutout or Random Erasing that remove random spatial regions, PartDrop leverages semantic priors to simulate realistic structural defects (e.g., a missing wing or wheel).

The process, formally expressed as $P_{partial} = \text{FPS}(\mathcal{G}_{mask}(P_{gt}, \mathcal{F}_{seg}(P_{gt})))$, involves three steps:

- **Semantic Decomposition:** We first employ a pre-trained PointNet++ segmenter $\mathcal{F}_{seg}$ [17] to decompose the ground truth P_{gt} into semantic parts $\mathcal{S} = \{S_1, \ldots, S_K\}$ based on ShapeNet taxonomies.
- **Stochastic Structural Masking:** We rank parts by size and iteratively select a subset of parts to "remove" by setting their coordinates to the origin $(0, 0, 0)$. This continues until a target removal ratio (50-75%) is met, generating a masked point cloud P_{masked} with broken semantic integrity.
- **Normalization via FPS:** Finally, Farthest Point Sampling (FPS) is applied to P_{masked}. Since the removed points are collapsed to a single singularity at the origin, FPS inherently filters them out by selecting diverse points from the remaining valid geometry, standardizing the output size to N_p.

By exposing the model to such structurally varied inputs, PartDrop enforces the learning of robust part-to-whole dependencies, significantly narrowing the domain gap between synthetic training and real-world occlusions.

B. Structure-Aware Dual-Stream Encoder

To robustly recover the underlying shape from partial observations while preserving fine-grained details, we propose a **Structure-Aware Dual-Stream Encoder**. As illustrated in Stage 1 of Figure 2, this module processes the partial input

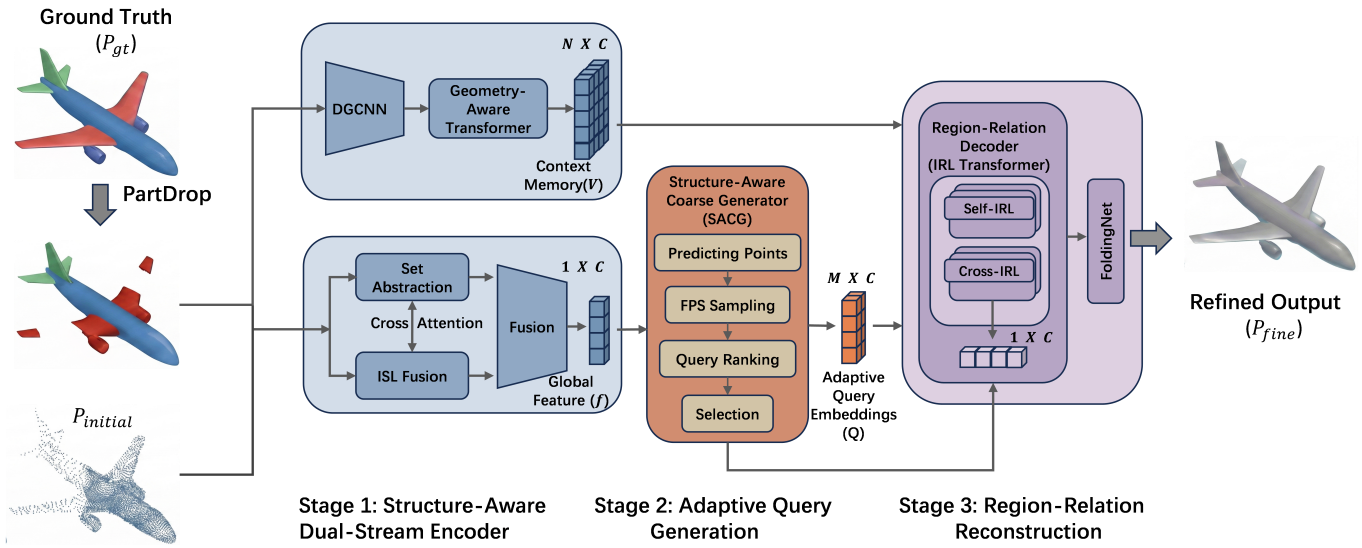

Fig. 2: The overall architecture of our proposed framework. The pipeline begins with **PartDrop**, a semantic-guided augmentation strategy that synthesizes structurally diverse partial inputs (P_{seg}) from ground truth shapes. The network then processes the input in three stages: **Stage 1 (Structure-Aware Dual-Stream Encoder)** extracts features via two interacting pathways: a Context Stream (top) generating Context Memory (V) via Geometry-Aware Transformers, and a Structure Stream (bottom) producing a Global Structure Feature (f) through Set Abstraction and ISL Fusion. **Stage 2 (Adaptive Query Generation)** employs the **Structure-Aware Coarse Generator (SACG)** to predict coarse geometry and dynamically select high-confidence anchors, forming Adaptive Query Embeddings (Q). **Stage 3 (Region-Relation Reconstruction)** utilizes a **Region-Relation Decoder** composed of Self-IRL and Cross-IRL Transformer blocks to model long-range structural dependencies, followed by FoldingNet to generate the final high-fidelity output (P_{fine}).

$P_{partial} \in \mathbb{R}^{N\times 3}$ through two parallel pathways—a *Context Stream* and a *Structure Stream*—to extract both dense point-wise features and a holistic structural descriptor.

a) Context Stream: Geometry-Aware Feature Extraction: The upper branch (Context Stream) is designed to generate a rich feature sequence that preserves local geometric details for subsequent decoding. Following the encoder design of PoinTr [1], we first employ a DGCNN [14] to extract local geometric features from the input point cloud. These features are then augmented with positional embeddings and processed by a **Geometry-Aware Transformer** encoder. Unlike standard Transformers that solely rely on feature similarity, this module explicitly incorporates a parallel geometric branch. It utilizes a k-NN operation on spatial coordinates to cluster local neighborhoods, aggregating their features via Max-Pooling to capture topological inductive bias. These local descriptors are fused with global semantic features derived from self-attention, producing the **Context Memory** $V \in \mathbb{R}^{N\times C}$, which serves as the key-value memory for the decoder's cross-attention layers.

b) Structure Stream: Hybrid Semantic-Topological Fusion: The lower branch (Structure Stream) aims to distill a compact global vector $\mathbf{f} \in \mathbb{R}^{1\times C}$ that encapsulates the overall semantic skeleton and topological integrity of the object. To achieve this, we design a hybrid block that synergistically fuses two distinct feature representations: **Set Abstraction (SA)** [17] for semantic context and **Intra-region Structure Learning (ISL)** [18] for local topology.

Specifically, the ISL module is engineered to capture fine-grained local geometry. As illustrated in Figure 3, for a point p_i with feature F_{p_i}, the ISL block extracts two distinct cues:

- **Neighbor-based Feature (T'):** This branch captures the local geometric topology relative to the k-nearest neighbors. It computes the feature difference between the center point and its neighbors, processes it with an MLP, and aggregates it via MaxPooling:

$$T'_{p_i} = \underset{j=1\ldots k}{\text{MaxPool}} \left(\text{MLP}_1(F_{p_i} - F_{p_{i,j}})\right). \quad (1)$$

- **Self-based Feature (T''):** This branch preserves absolute spatial information to prevent the loss of global pose context:

$$T''_{p_i} = \text{MLP}_2(F_{p_i}). \quad (2)$$

To optimally balance these cues—e.g., relying more on absolute pose when local topology is ambiguous due to occlusion—the block employs a learned gating mechanism (Sigmoid) to adaptively fuse them, producing the structure-enhanced feature T:

$$w = \sigma(\Phi(T' + T'')), \quad T = w \odot T' + (1 - w) \odot T'', \quad (3)$$

where Φ is a mixing MLP, σ is the Sigmoid function, and $\odot$ denotes element-wise multiplication.

To facilitate information flow between the SA and ISL representations, we introduce an internal **Cross-Attention** mechanism where the SA features query the ISL features (and vice versa) to mutually refine semantic and structural cues. Finally, the refined features from both branches are aggregated via a **Fusion** layer to produce the **Global Structure Feature** $\mathbf{f}$. This vector $\mathbf{f}$ provides the holistic guidance necessary for the subsequent *Structure-Aware Coarse Generator (SACG)*.

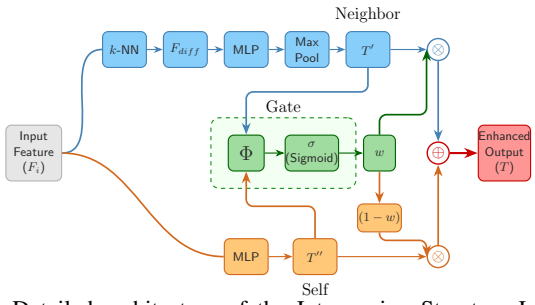

Fig. 3: Detailed architecture of the Intra-region Structure Learning (ISL) block. It extracts local geometric details via a Neighbor branch (using k-NN and MaxPool) and global pose information via a Self branch. A gating mechanism dynamically fuses these cues to produce the enhanced feature T.

C. Adaptive Query Generation

To enable high-fidelity reconstruction, we introduce an adaptive mechanism that generates dynamic queries tailored to the specific structural defects of the input. As illustrated in Stage 2 of Figure 2, this module, termed the **Structure-Aware Coarse Generator (SACG)**, utilizes the global structure feature $\mathbf{f}$ to regress high-quality structural anchors, which are then refined into adaptive query embeddings Q.

The generation process begins by establishing a comprehensive candidate set of structural anchors. We first employ a Multi-Layer Perceptron (MLP) to project the global feature $\mathbf{f}$ into a set of predicted coarse points $P_{pred} \in \mathbb{R}^{M\times 3}$, which primarily aim to fill in the missing regions based on learned structural priors:

$$P_{pred} = \text{MLP}_{gen}(\mathbf{f}). \tag{4}$$

To ensure the queries also faithfully preserve the observed geometry, we concurrently sample a subset of points P_{fps} from the original partial input using Farthest Point Sampling (FPS). These two sets are concatenated to form a robust candidate pool $P_{cand} = P_{pred} \cup P_{fps}$, which covers both inferred and observed structures.

However, simply using all candidate points may introduce redundancy or noise. To distill this set into high-quality anchors, we employ a learnable **Query Ranking** module. This module consists of a point-wise MLP that maps the 3D coordinates of each candidate point to a scalar significance score S, normalized by a Sigmoid activation:

$$S = \sigma(\text{MLP}_{rank}(P_{cand})), \tag{5}$$

where $P_{cand} \in \mathbb{R}^{(M+N/2)\times 3}$ and $S \in \mathbb{R}^{(M+N/2)\times 1}$. Based on these scores, we perform a sorting operation to rank all candidate points in descending order and select the top-N_q points to form the final coarse anchor set P_{coarse}. This selection strategy effectively filters out low-confidence predictions, ensuring that subsequent processing focuses on the most geometrically significant locations.

Finally, to generate the query embeddings Q for the decoder, we fuse the explicit positional priors of the selected anchors with the implicit global semantic guidance. Specifically, we expand the global feature $\mathbf{f}$ to match the number of anchors, concatenate it with the coordinates of P_{coarse}, and process the combined features through a shared MLP:

$$Q = \text{MLP}_{query}(\text{Concat}(\mathbf{f}, P_{coarse})). \tag{6}$$

These adaptive queries Q provide a robust initialization for the subsequent region-relation reconstruction, carrying both precise location information and rich structural context.

D. Region-Relation Reconstruction

The final stage of our pipeline is to translate the adaptive query embeddings $Q \in \mathbb{R}^{M\times C}$ into a high-fidelity, complete point cloud. As illustrated in Stage 3 of Figure 2, we design a sophisticated decoder that progressively refines these queries. Crucially, this stage leverages two key inputs from the encoding phase: 1) the **Context Memory** ($V \in \mathbb{R}^{N\times C}$) containing semantic features, and 2) the **Context Anchors** ($P_{ctx} \in \mathbb{R}^{N\times 3}$), which are downsampled coordinates from the original partial input. These inputs guide the reconstruction via a **Region-Relation Decoder (IRL Transformer)**.

a) Inter-region Relation Learning (IRL) Mechanism: To explicitly model structural dependencies while maintaining computational efficiency, we adapt the IRL unit [18] into a generative decoding block. Unlike standard attention, our IRL module operates on semantic regions. As detailed in Figure 4, the feature refinement process consists of three distinct phases:

1) Dynamic Region Construction: We decompose the feature space into S local regions. For the query stream, we use the coarse anchors P_{coarse} to select region centers via FPS. For the context stream (in Cross-IRL), we use the **Context Anchors** (P_{ctx}) to select centers. For each center, we utilize k-NN to group m neighboring points as regional representatives, producing a region tensor $\mathbf{X}_{region} \in \mathbb{R}^{S\times m\times C}$.

2) Grouped Relation Modeling: We treat the m representative points within each region as distinct "structural views" and perform **Grouped Multi-Head Attention**. The interaction is computed in parallel for each representative group index $g \in \{1, \dots, m\}$. Specifically, for the g-th group, we construct query matrices $\mathbf{Q}_g$, key matrices $\mathbf{K}_g$, and value matrices $\mathbf{V}_g$ of size $S \times C$. The inter-region affinity matrix $\mathbf{A}_g \in \mathbb{R}^{S\times S}$ is computed as:

$$\mathbf{A}_g = \text{Softmax}\left(\frac{\mathbf{Q}_g\mathbf{K}_g^\top}{\sqrt{C}}\right), \quad \mathbf{Y}_g = \mathbf{A}_g\mathbf{V}_g. \tag{7}$$

The outputs from all m groups are concatenated to form the enhanced region features $\mathbf{Y}_{region}$.

3) Feature Reconstruction via Interpolation: To map the region-level features back to the original resolution for residual connection, we employ an inverse distance-weighted interpolation based on the 3 nearest region centers:

$$Q_{out}^{(i)} = Q_{in}^{(i)} + \text{MLP}\left(\sum_{j\in\mathcal{N}_3(i)} w_{ij} \cdot \text{Mean}(\mathbf{Y}_{region}^{(j)})\right). \tag{8}$$

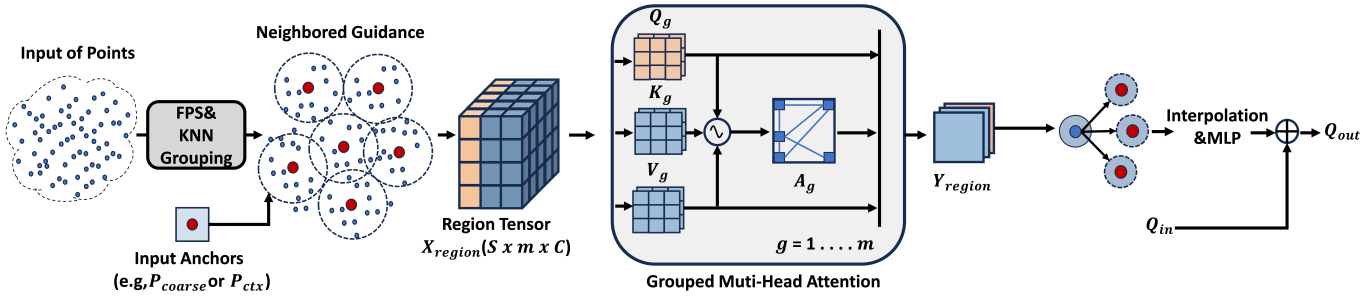

Fig. 4: Detailed illustration of the Inter-region Relation Learning (IRL) mechanism. It consists of three phases: 1) **Dynamic Region Construction** decomposes the input into local regions using FPS and k-NN; 2) **Grouped Relation Modeling** performs multi-head attention on grouped representatives to capture inter-region dependencies; 3) **Feature Reconstruction** maps the enhanced region features back to the original resolution via interpolation.

b) Cascaded Decoding Process: Our decoder employs a two-step interaction strategy:

- **Self-IRL (Query-to-Query):** First, we model dependencies among the generated parts. Here, both the queries and keys/values are derived from the input Q, utilizing P_{coarse} for region construction. This produces self-enhanced queries Q_{self}.
- **Cross-IRL (Query-to-Context):** Next, we align the generated parts with observed details. In this step, $\mathbf{Q}_g$ is derived from Q_{self} (using P_{coarse} for grouping), while $\mathbf{K}_g$ and $\mathbf{V}_g$ are derived from the **Context Memory** V (using **Context Anchors** P_{ctx} for grouping). This ensures that the generated queries can effectively attend to the spatially aligned features of the partial input, integrating global semantic context.

c) Geometry-Aware Decoding and Folding: Following the IRL Transformer, the refined queries are further processed by a standard Geometry-Aware Transformer Decoder [1] (using P_{ctx} as positional reference for keys) to capture fine-grained local geometric relationships. Finally, a FoldingNet-style [12] head generates the dense surface patches P_i around each anchor c_i:

$$P_i = f_{fold}(H_i) + c_i. \tag{9}$$

The final complete point cloud P_{fine} is formed by concatenating these generated patches.

IV. Experiments

A. Experimental Setup

To rigorously evaluate our method's capacity for learning generalizable shape priors, we conduct experiments on prominent benchmarks: ShapeNet-55 [19], ShapeNet-34, PCN [3], and KITTI.

a) Datasets and Protocols: **ShapeNet-55** assesses performance across diverse structures (55 categories). **ShapeNet-34** evaluates zero-shot generalization (training on 34 seen, testing on 21 unseen categories). The standard **PCN** dataset serves for targeted ablation studies. Crucially, to assess robustness against severe structural damage, we introduce the **Part-PCN** benchmark. It is constructed by applying our PartDrop strategy to PCN ground truth shapes, removing 75% of points starting from the largest semantic part. Finally, **KITTI** validates real-world performance.

b) Implementation Details: Inputs are standardized to 2,048 points. Training utilizes on-the-fly partial generation (removing 25%-75% points). We report Chamfer Distance (CD-$L1$, CD-$L2$) and F-Score@1% following standard protocols.

B. Results on Existing Benchmarks

1) Performance on PCN Dataset: Table I compares **SRCNet** against state-of-the-art approaches. Our method achieves the lowest average Chamfer Distance (6.48), outperforming Transformer-based competitors like PoinTr and AdaPoinTr. This demonstrates SRCNet's architectural superiority in capturing complex geometric structures. Visual comparisons are provided in Figure 5.

2) Performance on KITTI Dataset: We evaluated real-world generalization on KITTI by fine-tuning models pre-trained on ShapeNet cars. As shown in Table II, **SRCNet** achieves state-of-the-art Fidelity (0.000) and MMD (0.402), significantly outperforming PoinTr. This confirms our model's robustness in handling sparse, noisy LiDAR scans (qualitative results in Fig. 6).

C. Robustness to Structural Defects

To validate the hypothesis that viewpoint-based training fails to generalize to structural defects, we analyze performance on the **Part-PCN** benchmark.

Vulnerability of Conventional Models: As detailed in the "Std" (Standard) columns of Table III, state-of-the-art

TABLE I: Quantitative comparison on PCN. We report CD-L_1 ($\times 1000$) and F-Score@1%. Best results are **bolded**.

Method	Air.	Cab.	Car	Cha.	Lamp	Sofa	Tab.	Wat.	Avg	F1
PCN [3]	5.50	22.70	10.63	8.70	11.00	11.34	11.68	8.59	9.64	0.695
TopNet [20]	7.61	13.31	10.90	13.82	14.44	14.78	11.22	11.12	12.15	0.503
GRNet [21]	6.45	10.37	9.45	9.41	7.96	10.51	8.44	8.04	8.83	0.708
PoinTr [1]	4.75	10.47	8.68	9.39	7.75	10.93	7.78	7.29	8.38	0.745
SeedFormer [2]	3.85	9.05	8.06	7.06	5.21	8.85	6.05	5.85	6.74	–
AnchorFormer [22]	3.70	8.94	7.57	7.05	5.21	8.40	6.03	5.81	6.59	–
PointAttn [15]	3.87	9.00	7.63	7.43	5.90	8.68	6.32	6.09	6.86	–
PCDreamer [16]	**3.51**	**8.62**	**6.92**	6.91	5.66	8.31	6.27	5.90	6.52	**0.856**
SRCNet (Ours)	3.59	8.86	7.50	**6.73**	**5.14**	**8.43**	**5.94**	**5.67**	**6.48**	0.848

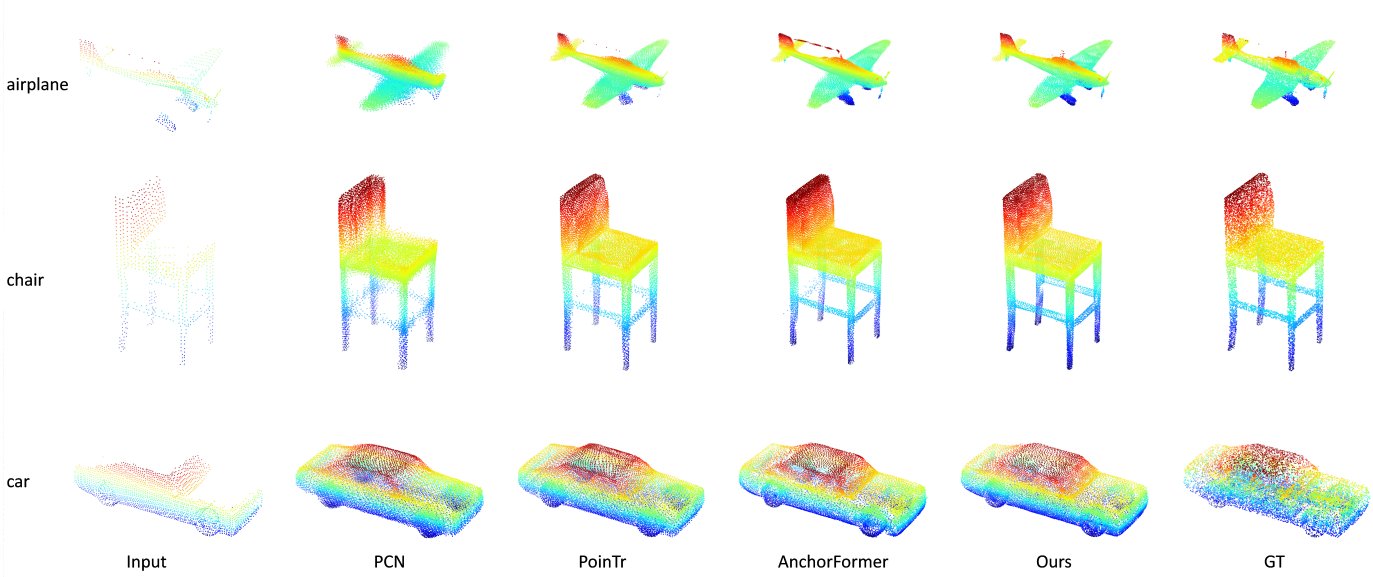

Fig. 5: Visual comparison of point cloud completion results on the PCN dataset.

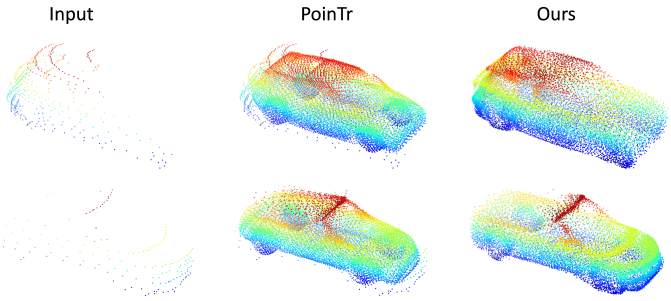

Fig. 6: Qualitative comparison on KITTI. Our SRCNet reconstructs fine details and coherent shapes from sparse LiDAR scans.

models exhibit dramatic performance degradation on Part-PCN compared to the standard PCN test set (e.g., PoinTr CD rises from 7.26 to 10.18). This confirms that models trained on simple defects learn shallow, view-specific priors rather than robust geometric reasoning.

Efficacy of PartDrop and SRCNet: We further disentangle the benefits of our contributions in Table III:

TABLE II: Quantitative comparison on the KITTI dataset. We report Fidelity and MMD ($CD{-}l_2 \times 1000$). Lower is better. The transposed format highlights the consistent superiority of SRCNet across both metrics.

Method	Fidelity ↓	MMD ↓
PCN [3]	1.76	2.11
FoldingNet [12]	2.24	1.37
TopNet [20]	7.47	0.54
GRNet [21]	1.02	0.87
SeedFormer [2]	0.82	0.57
PoinTr [1]	0.00	0.53
AnchorFormer [22]	0.00	0.46
SRCNet (Ours)	**0.00**	**0.40**

- **Impact of PartDrop:** Applying PartDrop (+PD) universally improves performance across all baselines. The gains are most significant on Part-PCN (e.g., reducing PoinTr's error by ∼15%), validating PartDrop as a robust augmentation strategy.
- **Architectural Superiority: SRCNet** trained with PartDrop outperforms all other combinations (CD_{L1} of 7.72 on Part-PCN). Notably, even when baseline models are enhanced with PartDrop, SRCNet remains superior, indicating that our dual-stream architecture is inherently better suited for reasoning about structural incompleteness.

D. Experiments on ShapeNet Benchmarks

To assess intrinsic architectural capabilities, we evaluate **SRCNet** on ShapeNet-55 and ShapeNet-34 *without* PartDrop augmentation, using standard training protocols.

ShapeNet-55 (Diversity): Table IV shows that SRCNet establishes a new state-of-the-art. It achieves the lowest average CD (0.82) and highest F-Score (0.538). Most importantly, in the **"Hard"** setting—which most closely mimics structural incompleteness—SRCNet significantly outperforms the Transformer-based SeedFormer (1.26 vs. 1.49).

TABLE III: Impact of training strategies. We compare **Std** (Standard Training) vs. **+PD** (training with PartDrop Augmentation). Evaluation is performed on the standard PCN test set and our challenging Part-PCN benchmark. Metrics are CD-L_1 ($\times 1000$).

Method	PCN (Easy)		Part-PCN (Hard)	
	Std	+PD	Std	+PD
PCN [3]	9.64	9.59	12.71	10.54
PoinTr [1]	7.26	7.22	10.18	8.59
AdaPoinTr [4]	6.53	6.50	9.26	7.86
SRCNet	6.51	**6.48**	9.21	**7.72**

TABLE IV: Quantitative results on ShapeNet-55. We report CD-L_2 ($\times$1000) and F-Score@1%. The column headers **S**, **M**, and **H** denote **S**imple, **M**oderate, and **H**ard difficulty levels, respectively.

Method	S	M	H	Avg	F1
FoldingNet [12]	2.67	2.66	4.05	3.12	0.082
PCN [3]	1.94	1.96	4.08	2.66	0.133
PoinTr [1]	0.58	0.88	1.79	1.09	0.464
SeedFormer [2]	0.50	0.77	1.49	0.92	0.472
PCDreamer [16]	-	-	-	0.91	0.408
SRCNet	**0.50**	**0.70**	**1.26**	**0.82**	**0.538**

ShapeNet-34 (Zero-shot Generalization): Table V evaluates performance on 21 unseen categories. SRCNet outperforms baselines on both seen (CD-Avg 0.73) and unseen (CD-Avg 1.17) categories. The minimal performance drop when transferring to novel classes suggests that our architecture learns transferable geometric priors rather than overfitting to category-specific training templates.

E. Ablation Study

To verify the effectiveness of each component in SRCNet, we conducted a progressive ablation study on the PCN dataset (Table VI).

Step-by-step Architectural Benefits: We started with a baseline model (Model A) where both the ISL encoder and IRL decoder modules were replaced by standard MLPs, yielding a CD of 6.77.

- **Effect of ISL:** Integrating the Intra-region Structure Learning (ISL) module (Model B) improves performance to 6.68. This indicates that explicitly capturing local topological details in the encoder is more effective than simple point-wise feature extraction.
- **Effect of IRL:** Further incorporating the Inter-region Relation Learning (IRL) decoder (Model C) significantly reduces the error to 6.51. This substantial gain (Δ0.17) confirms that modeling long-range dependencies between regions is critical for recovering complete shapes.

TABLE V: Zero-shot generalization on ShapeNet-34. We evaluate CD-L_2 ($\times$1000) on 34 **Seen** categories (Top) and 21 **Unseen** categories (Bottom). Columns S, M, and H represent Simple, Moderate, and Hard difficulties.

(a) Testing on 34 Seen Categories					
Method	S	M	H	Avg	F1
FoldingNet [12]	1.86	1.81	3.38	2.35	0.139
PCN [3]	1.87	1.81	2.97	2.22	0.154
PoinTr [1]	0.76	1.05	1.88	1.23	0.421
SRCNet	**0.43**	**0.64**	**1.13**	**0.73**	**0.483**

(b) Testing on 21 Unseen Categories					
Method	S	M	H	Avg	F1
FoldingNet [12]	2.76	2.74	5.36	3.62	0.095
PCN [3]	3.17	3.08	5.29	3.85	0.101
PoinTr [1]	1.04	1.67	3.44	2.05	0.384
SRCNet	0.59	0.93	1.99	1.17	**0.412**

TABLE VI: Step-wise ablation study. We progressively add **ISL** and **IRL** components (unchecked = replaced by MLP), and then compare augmentation strategies. Note that Random Drop degrades performance, while PartDrop yields the best result.

Model	ISL	IRL	Augmentation	CD-L_1
A			-	6.77
B	✓		-	6.68
C	✓	✓	-	6.51
D	✓	✓	Random Drop	6.59
E (Ours)	✓	✓	**PartDrop**	**6.48**

PartDrop vs. Random Drop: Finally, we evaluated data augmentation strategies based on the full architecture (Model C). Naively applying "Random Drop" (Model D) degraded performance to 6.59, suggesting that breaking geometry without semantic logic introduces harmful noise. In contrast, our **PartDrop** strategy (Model E) successfully boosts performance to 6.48. This demonstrates that *semantic-guided* occlusion forces the network to learn robust structural priors, whereas blind deletion hinders training.

V. Discussion

While SRCNet demonstrates superior robustness, we acknowledge specific limitations and trade-offs.

a) Dependency on Segmentation Granularity: PartDrop relies on the quality of the upstream segmenter. Current pre-trained models often lack fine-grained partitioning, limiting the diversity of synthesized defects. Moreover, for structurally simple objects, removing a dominant part can render the remaining geometry semantically ambiguous, potentially introducing noise during training. Future work could leverage open-vocabulary segmentation to generate richer and more precise structural masks.

b) Computational Efficiency: Our Structure-Aware Dual-Stream design explicitly trades computational cost for robust structural reasoning. As shown in Table VII, SRCNet operates at **17.95 GFLOPs**. While this incurs a moderate overhead compared to lightweight baselines like PoinTr (10.41 G), it remains comparable to SnowflakeNet (17.16 G) and is significantly more efficient than heavy architectures like GRNet (40.44 G). We consider this a justifiable trade-off for the substantial gains in completion fidelity and robustness.

VI. Conclusion

In this work, we identify and address the "structural blindness" of existing completion methods, which struggle with non-uniform, real-world defects. We propose a synergistic framework to enforce structural invariance through both data and architecture. First, we introduce **PartDrop**, a semantic-guided augmentation strategy that forces the network to learn robust part-to-whole dependencies rather than shallow pattern matching. Second, we propose the **Structural-Relational Completion Network (SRC-Net)**. By integrating a Structure-Aware Dual-Stream Encoder with an Inter-region Relation

TABLE VII: Complexity analysis. We compare the theoretical computation cost (FLOPs) of SRCNet against state-of-the-art methods. SRCNet maintains a balanced trade-off between complexity and performance.

Method	FLOPs (G)
TopNet [20]	6.72
PoinTr [1]	10.41
AdaPoinTr [4]	12.43
SeedFormer [2]	13.97
PCN [3]	15.25
SnowflakeNet [23]	17.16
GRNet [21]	40.44
SRCNet (Ours)	17.95

Learning (IRL) mechanism, SRC-Net explicitly models long-range dependencies between disjointed regions. Extensive experiments on PCN, ShapeNet-55, and KITTI benchmarks demonstrate that our approach achieves state-of-the-art performance, exhibiting superior robustness and generalization against severe structural incompleteness.

REFERENCES

[1] X. Yu, Y. Rao, Z. Wang, Z. Liu, J. Lu, and J. Zhou, "Pointr: Diverse point cloud completion with geometry-aware transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 498–12 507.

[2] H. Zhou, Y. Cao, W. Chu, J. Zhu, T. Lu, Y. Tai, and C. Wang, "Seedformer: Patch seeds based point cloud completion with upsample transformer," in *European conference on computer vision*. Springer, 2022, pp. 416–432.

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

[4] X. Yu, Y. Rao, Z. Wang, J. Lu, and J. Zhou, "Adapointr: Diverse point cloud completion with adaptive geometry-aware transformers," 2023. [Online]. Available: https://arxiv.org/abs/2301.04545

[5] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, vol. 7, 2006.

[6] J. Davis, S. R. Marschner, M. Garr, and M. Levoy, "Filling holes in complex surfaces using volumetric diffusion," in *First International Symposium on 3D Data Processing Visualization and Transmission, 2002. Proceedings.* IEEE, 2002, pp. 428–438.

[7] N. J. Mitra, L. J. Guibas, and M. Pauly, "Partial and approximate symmetry detection for 3d geometry," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 560–568, 2006.

[8] M. Pauly, N. J. Mitra, J. Wallner, H. Pottmann, and L. Guibas, "Discovering structural regularity in 3d geometry," in *ACM SIGGRAPH 2008 papers*, 2008, pp. 1–11.

[9] C.-H. Shen, H. Fu, K. Chen, and S.-M. Hu, "Structure recovery by part assembly," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 6, pp. 1–11, 2012.

[10] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

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

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

[13] M. Sarmad, H. J. Lee, and Y. M. Kim, "Rl-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5898–5907.

[14] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[15] J. Wang, Y. Cui, D. Guo, J. Li, Q. Liu, and C. Shen, "Pointattn: You only need attention for point cloud completion," in *Proceedings of the AAAI Conference on artificial intelligence*, vol. 38, no. 6, 2024, pp. 5472–5480.

[16] G. Wei, Y. Feng, L. Ma, C. Wang, Y. Zhou, and C. Li, "Pcdreamer: Point cloud completion through multi-view diffusion priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 243–27 253.

[17] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[18] S. Cheng, X. Chen, X. He, Z. Liu, and X. Bai, "Pra-net: Point relation-aware network for 3d point cloud analysis," *IEEE Transactions on Image Processing*, vol. 30, pp. 4436–4448, 2021.

[19] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

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

[21] H. Xie, H. Yao, S. Zhou, J. Mao, S. Zhang, and W. Sun, "Grnet: Gridding residual network for dense point cloud completion," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 365–381.

[22] Z. Chen, F. Long, Z. Qiu, T. Yao, W. Zhou, J. Luo, and T. Mei, "Anchorformer: Point cloud completion from discriminative nodes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 13 581–13 590.

[23] P. Xiang, X. Wen, Y.-S. Liu, Y.-P. Cao, P. Wan, W. Zheng, and Z. Han, "Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5479–5489.