A Point Cloud Completion Network Via The Latent Space-driven Two-stage Noise Synthesis And Restoration Strategy

1st Xiaofei Qin*, 2nd Anluo Yi*, 3rd Shiwei Tao*, 4th Jie Zhang*, 5th Xuedian Zhang* †
*School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
†Corresponding Author: obmmd_zxd@163.com
Other Emails: xiaofei.qin@usst.edu.cn, {232260517, 243370925, 233370860}@st.usst.edu.cn

***Abstract*—Raw point cloud data collected in real-world scenarios often encounter complex interferences such as sensor noise and uneven density, making high-fidelity point cloud completion tasks extremely challenging. Current mainstream point cloud completion methods generally suffer from insufficient noise resistance, struggling to meet practical application requirements. Moreover, publicly available datasets for noisy point cloud completion are relatively scarce. We proposes a point cloud completion network via the latent space-driven two-stage noise synthesis and restoration strategy, constructing a novel unified framework. Within this framework, noise synthesis and restoration are achieved through gradient guidance and constrained projection mechanisms under the constraint of hyperspherical distribution in the feature-level latent space. Experimental results on public datasets demonstrate that our method significantly improves both noise robustness and completion accuracy compared to the current state-of-the-art (SOTA) techniques, offering a new solution for noisy point cloud completion tasks.**

***Index Terms*—Point cloud completion, latent space, gradient affine transformation, noise synthesis and restoration**

I. Introduction

Point cloud completion aims to rectify occlusions, sampling constraints, or sparsity-induced missing points, providing high-quality inputs for downstream tasks [1], [2]. Its core lies in learning implicit 3D distributions to transition from partial observations to complete geometric models [3]–[6]. As depicted in the first line of Fig.1, typical models adhere to a two-stage *feature extraction - point cloud reconstruction* paradigm. In the feature extraction phase, the model extracts global shape features from unordered, irregular point clouds. Given the permutation invariance of point clouds (i.e., point order doesn't alter geometric representation), symmetric functions (e.g., max/average pooling) are employed for feature extraction. Based on the different main networks used by $f_{\mathcal{T}(\cdot)}, f_{\mathcal{A}(\cdot)}, f_{\mathcal{R}(\cdot)}$, the existing point cloud completion methods can be roughly divided into two categories: those based on convolutional methods [3], [4], [7] and those based on Transformer methods [5], [6], [8].

Recently, Transformer-based point completion methods have garnered significant attention due to their exceptional global modeling capabilities, as their attention mechanisms enable precise modeling of global point cloud dependencies and learning of discriminative latent representations. In our prior work [9] (detailed in **Appendix**), we proposed SPA-Net (second line of Fig.1), which achieved state-of-the-art performance on ShapeNet-55/34 [10] and MVP [11] benchmarks. Within the Transformer-based point cloud completion algorithm paradigm, the Encoder's primary task is to transform input point cloud data into high-dimensional semantic feature representations, while the Decoder generates a complete point cloud based on these semantic features. Our previous work, along with most existing Transformer-based representative baselines such as PionTr [10], 3DMamba [12], and SeedFormer [8], incorporates a multi-scale neighboring feature aggregator before the Encoder to progressively extract local and global features from the point cloud, converting the input point cloud into point embeddings equivalent to projecting it into a high-dimensional latent semantic space where each dimension represents a certain semantic feature.

However, these methods, including ours, lack specific opti-

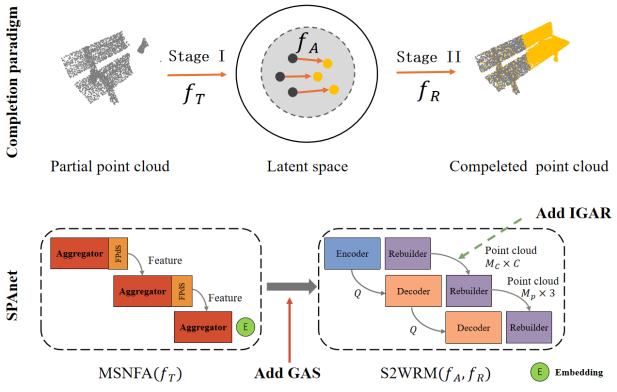

Fig. 1. Architecture of two-stage point cloud completion paradigm and SPA-net. **Paradigm** map known points $P_{know}$to latent space $Z_{konw} \subseteq \mathbb{R}^d$. In reconstruction, they get missing point cloud features $f_{\mathcal{A}}(Z_{\text{know}})$ from known ones in latent space and decode full 3D coordinates $f_{\mathcal{A}}(Z_{\text{know}})$. **SPA-Net** consists of MSNFA (aggregating point clouds with high info density in small dims) and S2WRM (making level-by-level predictions for multi-scale outputs$\{\mathcal{P}_1 \in \mathbb{R}^{M_c \times 3}, \mathcal{P}_2 \in \mathbb{R}^{M_p \times 3}, \mathcal{P}_3 \in \mathbb{R}^{M \times 3}\}, M : M_p : M_c = 16 : 4 : 1$.

This work was supported by the Shanghai Magnolia Talent Program Pujiang Project under [25PJD089].

mization for point cloud noise in practical scenarios, making them noise-sensitive and unable to fully meet deployment/downstream task demands, and the scarcity of realistic noisy public datasets further limits related research and applications. Moreover, the self-attention mechanism in the Encoder allows each point embedding to interact with all others, capturing their relationships and mutual influences, but when noise points are present, the multi-scale neighboring feature aggregator generates incorrect point embeddings, affecting subsequent completion. This constitutes the core motivation for the technical improvements in this paper, necessitating algorithm enhancements to address two main issues: (1) effectively identifying noise in real-world environments where multiple factors cause point cloud noise; (2) Due to systematic errors in devices, the point clouds of an object may all exhibit varying degrees of disturbance. Simply removing noise points may result in the loss of significant information.

In generative models [13]–[15], the latent space plays a pivotal role in noise suppression and anomaly detection. According to latent space theory, a sufficiently accurate fitting projection $f_{\mathcal{T}(\cdot)}$ yields distinct distributions: normal points of the same object adhere to a specific manifold, while noise points deviate as anomalies. Analogous to the scarcity of labeled noisy point cloud completion datasets, anomaly detection often operates under unsupervised/small-sample regimes, where latent space-driven methods [16], [17] leverage this discrepancy to generate synthetic negatives and mitigate data scarcity. Inspired by this, we propose a latent space-driven two-stage noise synthesis/restoration strategy, integrating it into SPA-Net to form SPAn-Net; this rarely explored strategy in prior point cloud completion research distinguishes noise from normal points via latent semantic embedding distribution differences, encompassing Gradient Affine Synthesis (GAS) and Iterative Gradient-Aware Restoration (IGAR) (second line of Fig.2). Our contributions are summarized as follows:

- We introduced the GAS stage, which, through adversarial learning, synthesizes diverse pseudo-noise in the absence of sufficient noisy data. This enables the precise fitting of the normal boundary in latent semantic space, facilitating accurate identification of noise points by distinguishing their distribution differences from normal points.
- We introduced the IGAR stage to map identified noise embeddings back inside the hypersphere, generating pseudo-embeddings with normal semantic features while retaining positional information to synergize with the Encoder's self-attention mechanism. We employed an oversaturation point generation strategy during initial reconstruction, coupled with a second discriminator for confidence screening, dynamically balancing feature preservation and noise suppression to enhance reconstruction robustness.
- We proposed SPAn-net, which builds upon our prior SPA-Net by integrating two novel stages—GAS and IGAR—to achieve state-of-the-art performance in point cloud completion on both the simulated noisy dataset ProjectShapeNet-55/34 and the real-world point cloud dataset KITTI.

II. Related work

A. Point cloud completion

Point cloud completion is a generative and estimation problem derived from partial point clouds, playing a crucial role in the applications of 3D computer vision. According to the used different backbones, the existing point cloud completion methods could be roughly divided into two categories: convolution-based methods and Transformer-based methods.

The existing convolution-based [3], [4], [7] generally handle the point cloud completion task by extracting features from the input incomplete point clouds via either 3D or 2D convolutions. However, the convolution-based methods, are essentially based on Convolutional Neural Networks (CNNs). These networks are characterized by limited receptive fields and struggle with modeling variable-size inputs and long-range semantic relationships [18]. Consequently, they often exhibit suboptimal performance in point cloud completion tasks.

To address this problem, many Transformer-based methods [5], [6], [8], [19] have been proposed recently, leveraging the capability of Transformers to effectively manage long-range dependencies in point clouds [18]. PionTr [10], presents an adaptive point cloud Transformer framework for point cloud generation. These two works demonstrate that Transformers are proficient in modeling long-range dependencies between local structures within point clouds. Subsequently, many Transformer-based methods have emerged [5], [20]–[22]. However, as these methods have noted, Transformers inherently lack geometric priors (such as translation invariance) that are characteristic of CNNs. This deficiency may result in the loss of structural knowledge or detailed local information, thereby making it challenging to recover these details during decoding. Consequently, the contributions of previous Transformer-based methods have primarily focused on developing mechanisms to extract local and structural information from point cloud data. For example, PoinTr [10] devises a geometry-aware block that models the local geometric relationships explicitly. SeedFormer [8] introduces a novel shape representation named Patch Seed and develops an enhanced upsample Transformer for the generation process. ProxyFormer [23] strengthens structural knowledge through a proxy alignment mechanism. GSFormer [19] proposes a novel graph-structured representation for point cloud generation.

B. Noise processing based on latent space

In generative models, the latent space, serving as a high-order abstract representation of data, has emerged as a pivotal vehicle for noise suppression and anomaly detection [13]–[15]. Through mechanisms such as manipulating the latent distribution of features, constraining reconstruction errors, or adversarial training, it has demonstrated remarkable efficacy in image, time-series, and graph-structured data. DSCNet [24] integrates latent representations with subspace regularization,

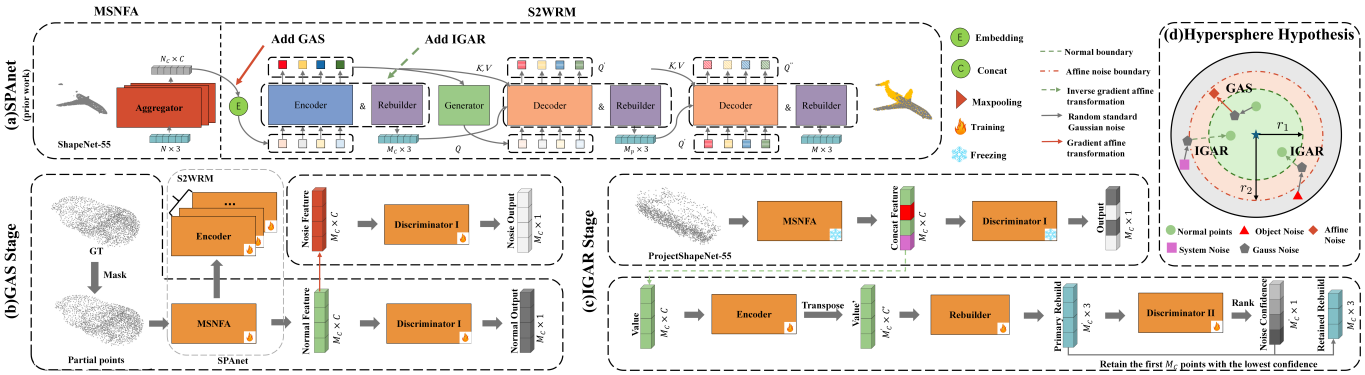

Fig. 2. Architecture of SPAn-Net. **(a) SPAn-Net** additional incorporates two sequentially trained stages: The **(b) GAS stage** then simulates missing data by randomly masking the ground truth point cloud from ProjectShapeNet-55, using MSNFA to process these partial points. It employs gradient affine transformations to create pseudo-noise points for training Discriminator I, which learns noise discrimination. In the **(c)IGAR stage**, the model inputs noisy, incomplete point clouds, freezes MSNFA and Discriminator I weights, and generates an oversaturated point cloud. Discriminator II filters out noise points, retaining only high-quality points. **(d)Hypersphere Hypothesis Diagram**, illustrates how normal points become affine noise (via GAS stage) and how system/object noise is restored to normal features (via IGAR stage) within the defined boundaries in the latent space.

achieving automatic pruning of noise subspaces through low-rank constraints, thereby enabling precise segmentation of tubular structures (e.g., blood vessels, roads). For time-series data, OmniAnomaly [25] employs stochastic variable connection techniques to model multimodal latent trajectories, achieving state-of-the-art (SOTA) performance in server log anomaly detection. ADGAN [26] utilizes latent space interpolation to generate synthetic anomalies, addressing the challenge of missing negative samples and thereby facilitating adversarial training of the model. GLASS [27] employs latent space mapping to controllably synthesize a broader range of industrial anomaly samples.

For point cloud data, traditional methods typically employ various filtering techniques, such as Gaussian filtering and mean filtering, to smooth noise. However, these approaches are ineffective in handling sparse point clouds captured in real-world scenarios. Latent space methods, by transforming noise identification and removal into a decoupling problem in high-dimensional feature spaces, have validated their effectiveness in the field of 2D images. Therefore, in this paper, noise identification and removal are also conducted within the latent space.

III. Method

A. Model architecture

According to latent space theory, if the multi-scale neighboring feature aggregator fits the projection process accurately enough, the positions of noise points in the latent semantic space will be entirely different from those of normal points. This is because all normal points originate from the same object, and their distribution in the latent semantic space follows certain patterns. Inspired by this, we utilizes the differences in semantic features between noise points and normal points for noise identification, distinguishing the two based on the differences in their distributions in the latent semantic space. Specifically, we assumed that the normal point cloud of the same object follows a hyperspherical model in the latent semantic space. As shown in Figure 3, all normal point cloud embeddings of each object in the latent semantic space are enveloped by a sphere, referred to as the "Normal boundary" in this paper.

To enhance the accuracy of the fitted normal boundary, we employ an adversarial learning strategy and devise a Gradient Affine Transformation in GAS stage. GAS, as a form of data augmentation, first addresses the issue of data scarcity by applying gradient affine transformations to normal point embeddings, thereby generating pseudo-noise point embeddings. This process synthesizes a wide array of training scenarios, enabling the model to learn how to recognize and handle various noise types effectively, even without relying on extensive real-world noisy data.

GAS stage provides a controllable method for synthesizing affine noise with real noise characteristics, enabling discriminator I $\mathcal{D}_1$ to output confidence scores $\epsilon_i \in [0, 1]$ for the noise embedded in each point, and to learn precise fitting of normal boundaries ($\hat{r} \rightarrow r_1$) through a similar adversarial training approach. Specifically, we first generate partial inputs by randomly removing n farthest points ($25\% \leq n \leq 75\%$) from ground-truth point clouds $\mathcal{G} \in \mathbb{R}^{8192\times 3}$ and then downsampling the 2048 remaining points. During GAS stage training, normal points embeddings $U = \{u_i \in \mathbb{R}^{1\times C}\}$ are mapped inside a hypersphere ($\mathcal{D}_1(u_i) \rightarrow 0$). Conversely, for each affine noise point embedding $v_i \in U_v$ (generated via gradient-based transformations introduced in III-B), the discriminator should output a confidence score of 1 , and satisfy $r_1 \leq \|v_i - c\|_2 \leq r_2$, where c denotes the hypersphere center. The discriminator $\mathcal{D}_1$ is trained to minimize:

$$\min_{\mathcal{D}_1} [\mathbb{E}_{u_i \sim U} \log \mathcal{D}_1(u_i) + \mathbb{E}_{v_i \sim U_v} \log(1 - \mathcal{D}_1(v_i))] \quad (1)$$

where $\mathbb{E}$ represents the expectation. Complementarily, IGAR offers a soft correction approach for noisy point embeddings. Given that systematic errors in devices can cause all point

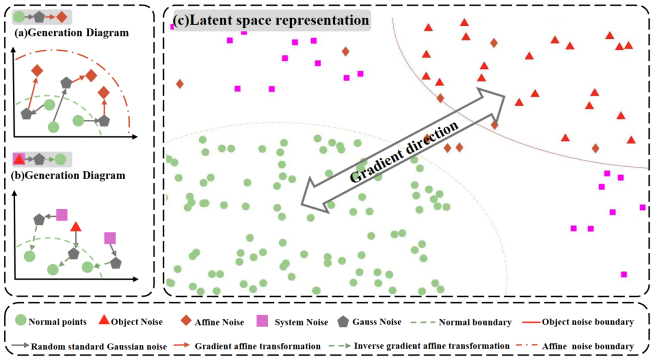

Fig. 3. Representation of latent space. (a) Gradient affine transformation. (b) Inverse gradient affine transformation. (c) Distribution of different point embeddedings

clouds of an object to suffer from varying degrees of disturbance, simply eliminating noise points often leads to the loss of crucial information. In contrast, IGAR preserves valuable geometric cues by retaining potentially useful 3D coordinates rather than outright discarding them. This enables it to restore the synthesized noisy points to their correct positions, ensuring that the final point cloud is not only accurate but also consistent with the expected data distribution, effectively addressing the challenge of information loss due to device-induced disturbances.

IGAR stage introduces a novel noise restoration paradigm that recovers the semantics of noisy points to normal semantics in latent space while preserving structural information. With frozen $\mathcal{D}_1$, the noisy input point cloud (from train data of noisy completion datasets) embeddings set $\mathcal{X}$ are first filtered by threshold τ:

$$\mathcal{X}_n = \{x_i \in \mathcal{X} \mid \mathcal{D}_1(x_i) \geq \tau\} \tag{2}$$

$x_i \in \mathcal{X}_n$ are then restored to the hypersphere interior ($\|u_i' - c\|_2 \leq r_1, u_i' = restored(x_i)$) via inverse gradient-affine transformations introduced in III-B. Additionally, we adopt an oversaturation point generation strategy during the $f_{\mathcal{R}}$. The encoder first sets embedding dimension $M_c' = (1+\alpha)M_c$ to produce $U_c' \in \mathbb{R}^{M_c' \times 3}$, then secondary discriminator $\mathcal{D}_2$ scores each point:

$$\epsilon_i = \mathcal{D}_2(p_i), \quad p_i \in U_c' \tag{3}$$

Finally, the top M_c points $\mathcal{P}_{refined} \in \mathbb{R}^{M_c \times 3}$ with lowest noise scores are selected through ranking operations.

B. Latent space representation and gradient affine transformation

According to latent space theory, normal samples $U = \{u_i \in \mathbb{R}^{1 \times C}\}$ adhere to the Hypersphere Hypothesis: their embeddings concentrate on or near a hyperspherical surface with center c and radius r_1. Noise embeddings are formally defined as:

$$U' = \{\tilde{u}_i \mid \|\tilde{u}_i - c\|_2 > r_1, \tilde{u}_i \in U\} \tag{4}$$

As illustrated in Fig.3, noise in real point cloud data within the latent space can be categorized into two types: (1)random system noise that completely deviates from the data distribution (Out-of-Distribution), and (2) object noise that approximately follows the data distribution (Near-in-Distribution). To simulate realistic noise, we propose a gradient-affine transformation which controls the generation direction and offset of these affine noises through gradient guidance and constrained projection, respectively.

Gradient guidance: The transformation begins with Gaussian noise injection, where random standard normal noise $\sigma_i \sim \mathcal{N}(0, I)$ is added to normal points embeddings: $g_i = u_i + \sigma_i$, with $I \in \mathbb{R}^{C \times C}$ denoting the identity matrix. This induces random latent space displacements, which means it cannot guarantee $g_i \in U'$. To address this, the optimization of the gradient guidance is performed by iteratively adjusting g_i along the discriminator loss gradient $\nabla \mathcal{L}_A$:

$$\tilde{g}_i = g_i + \eta \frac{\nabla \mathcal{L}_A(g_i)}{\|\nabla \mathcal{L}_A(g_i)\|} \tag{5}$$

where η is the learning rate. Normal samples typically incur lower losses while anomalous ones (noise in this paper) lead to higher losses. Iterative gradient guidance in the latent space thus shifts samples radially outward from the hyperspherical center, which makes $\tilde{g}_i$ resemble anomalous samples in terms of latent space characteristics.

Constrained projection : Additionally, the generated noise $\tilde{g}_i$ must inherently satisfy two geometric constraints: be located outside the hypersphere boundary of normal data($\|\tilde{g}_i - c\|_2 > r_1$); the maximum offset distance is constrained to prevent discriminator overfitting and blurring the fitting boundary (Avoid $\|\tilde{g}_i - c\|_2 >> r_1$). The offset vector $\theta_i = \tilde{g}_i - u_i$ undergoes threshold-based normalization:

$$\widehat{\theta}_i = \frac{\alpha_i}{\|\theta_i\|}\theta_i, \quad \alpha_i = \begin{cases} r_1 & \|\theta_i\| < r_1 \\ r_2 & \|\theta_i\| > r_2 \\ \|\theta_i\| & \text{otherwise} \end{cases} \tag{6}$$

where $r_2 = 2r_1$ defines the maximum allowable displacement. The final affine noise feature is computed as $v_i = u_i + \widehat{\theta}_i$. For noise restoration tasks, an inverse gradient-affine transformation is defined by shifting real noise features $\tilde{u}_i$ along the gradient descent direction: $u_i' = \tilde{u}_i - \widehat{\theta}_i$.

C. Loss and inference

During the GAS stage, the total loss L_C comprises three components: the backbone network (SPA-net) completion loss L_{CD} and two discriminator losses (L_A and L_N) from Discriminator I. SPA-net generates multi-scale predictions $\{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\}$. The Chamfer Distance (CD) L1 loss is computed for each scale and aggregated:

$$L_{CD} = \sum_{k=1}^{3} d_{CD}(\mathcal{P}_k, \mathcal{G}), \tag{7}$$

$$d_{CD}(\mathcal{P}, \mathcal{G}) = \frac{1}{|\mathcal{P}|}\sum_{p \in \mathcal{P}} \min_{g \in \mathcal{G}} \|p - g\| + \frac{1}{|\mathcal{G}|}\sum_{g \in \mathcal{G}} \min_{p \in \mathcal{P}} \|g - p\|. \tag{8}$$

The discriminator I outputs confidence scores $\mathbf{E}_n \in \mathbb{R}^{N\times 1}$ for normal samples and $\mathbf{E}_A \in \mathbb{R}^{N\times 1}$ for affine noise samples $U_v = \{u_i + \widehat{\theta}_i\}$. The losses are:

$$L_N = f_{BCE}(\mathbf{E}_n, \mathbf{0}), \tag{9}$$

$$L_A = f_{BCE}(\mathbf{E}_A, \mathbf{1}), \tag{10}$$

where $\mathbf{0}$ and $\mathbf{1}$ are zero and one vectors of dimensionality $N \times 1$. The total adversarial loss is:

$$L_C = L_{CD} + \beta(L_A + L_N), \tag{11}$$

where β balances the loss contributions. During IGAR stage, the aggregator and Discriminator I weights are frozen. Discriminator II refines the initial reconstruction $\mathcal{P}_1$ and is trained solely with L_{CD}:

$$L_{refine} = d_{CD}(\mathcal{P}_{refined}, \mathcal{G}). \tag{12}$$

IV. Experiment

A. Datasets

ProjectShapeNet-55 and ProjectShapeNet-34 [6] are alternative versions of the ShapeNet-55 and ShapeNet-34 [10], which more closely resemble real-world scenarios with missing point cloud data (characterized by noise and greater sparsity). These datasets share similar settings and possess identical ground truth labels.

In the ShapeNet-55 and ShapeNet-34, incomplete point clouds are generated by directly cropping complete point clouds. In contrast, ProjectShapeNet-55 and ProjectShapeNet-34 are generated through noisy back-projection of depth images from a single viewpoint. For each sample, 16 viewpoints are randomly selected to render depth images, and noisy back-projection is performed to obtain 16 different incomplete point clouds as training inputs. ProjectShapeNet-55 uses all object categories for both training and testing. In contrast, ProjectShapeNet-34 utilizes only 34 visible object categories for training and evaluates the model's performance on both visible and invisible categories.

KITTI [29] ranks among the most authoritative computer vision algorithm evaluation datasets in the autonomous driving domain. It primarily addresses scene tasks in the in-vehicle environment, including stereo vision, optical flow estimation, visual odometry, 3D object detection, and tracking. The data is collected from real-world urban, rural, and highway scenarios, encompassing dynamic objects (such as vehicles, walkers, and cyclists) under complex traffic conditions, along with multi-sensor fusion data. We evaluate the model's performance on the KITTI dataset, utilizing incomplete point clouds of cars in real-world scenarios scanned by LiDAR.

B. Evaluation Metrics

For the ProjectedShapeNet-55 and ProjectedShapeNet-34 datasets, the evaluation metrics are similar to those used for ShapeNet-55/34. Specifically, the L1 chamfer distance (see Eq. (8) for the calculation method of CD-ℓ_1) and F-score@1% are used for evaluation. We define the precision and recall of the point cloud completion results at threshold d (set to 1% by following previous works) as:

$$P(d) = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\left[\min_{g\in\mathcal{G}}\|p-g\| < d\right] \tag{13}$$

$$R(d) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\left[\min_{p\in\mathcal{P}}\|g-p\| < d\right] \tag{14}$$

$$F\text{-}score(d) = \frac{2P(d)\cdot R(d)}{P(d)+R(d)} \tag{15}$$

where $P(d)$ and $R(d)$ denote the precision and recall.

For the KITTI dataset, the evaluation metrics [5] include minimum matching distance (MMD) and fidelity. MMD represents the chamfer distance between the closest output point cloud and the ground truth point cloud, which measures the similarity between the former and the latter. Fidelity refers to the average distance from each point in the input point cloud to its nearest neighbor in the output point cloud. This metric is used to evaluate the extent to which the input information is preserved.

C. Implementation Detail

In the model training process, two NVIDIA RTX 3090 graphics cards are used to train the model separately on the ProjectedShapeNet-55 and ProjectedShapeNet-34 datasets for evaluation on these two datasets. For the KITTI dataset, a model pre-trained on ProjectedShapeNet-55 is fine-tuned on the KITTI dataset for 20 epochs for evaluation. The initial learning rate is set to 10^{-4}. An AdamW optimizer with a weight decay of 5×10^{-3} is selected. The weight coefficient is initially set to 10^{-2} and is multiplied by ten every 100 training epochs until it reaches 1. The total number of training epochs is 400. The initial number of point embeddings in the SPA-net encoder is $M_c = 512$. If the saturation rate a is set to $\alpha = 0.25$, then the number of point embeddings in the encoder is $512 \times (1 + 0.25) = 640$, $\beta = 0.5$.

D. Comparison results

Results on ProjectedShapeNet-55. As shown in Table I, the proposed method achieves a CD-ℓ_1-Avg of 8.85, outperforming all existing approaches. For eight representative object categories (table, airplane, car, sofa, birdcage, remote, keyboard, and rocket), the proposed method attains state-of-the-art performance in terms of CD-ℓ_1-Avg. Although it slightly underperforms AdaPoinTr on the F-score@1% metric, visual comparisons in Fig.4 demonstrate that, even when handling input point clouds with severe noise, the proposed method generates completion results that maintain high geometric consistency with ground truth.

Results on ProjectedShapeNet-34. Table II presents the test results on the ProjectShapeNet-34 dataset. The model is trained on 34 seen categories and is evaluated separately on both 34 seen and 21 unseen categories. Experimental results demonstrate that the proposed method significantly

TABLE I
COMPARISON ON THE PROJECTEDSHAPENET-55. CD-ℓ_1-AVG DENOTES THE AVERAGE CD-ℓ_1 $\times$ 10^3 ACROSS ALL CATEGORIES. AND ADDITIONAL RESULTS FOR CERTAIN CATEGORIES WERE DISPLAYED.

Method	Table	Airplane	Car	Sofa	Birdcage	Remote	Keyboard	Rocket	**CD-ℓ_1-Avg↓**	**F-score@1%↑**
GRNet [3]	12.01	8.30	12.13	14.36	16.52	12.18	9.71	8.58	12.81	0.491
PoinTr [10]	9.97	6.02	10.58	12.11	14.60	9.55	7.61	6.86	10.68	0.615
Snowflake [5]	10.49	6.35	11.20	12.59	15.24	10.07	8.12	7.49	11.34	0.594
AdaPoinTr [6]	8.81	5.18	9.77	10.89	13.27	8.81	6.79	5.58	9.58	**0.701**
SVDFormer [28]	9.15	5.45	10.12	11.25	13.65	9.18	7.15	5.95	9.71	0.689
Ours	**8.70**	**5.16**	**8.61**	**10.55**	**11.88**	**8.27**	**5.90**	**5.37**	**8.85**	0.689

TABLE II
COMPARISON ON THE PROJECTEDSHAPENET-34.

Method	**34 seen categories**		**21 unseen categories**	
	CD-ℓ_1-Avg↓	F-Score@1%↑	CD-ℓ_1-Avg↓	F-Score@1%↑
GRNet [3]	12.41	0.506	15.03	0.439
PoinTr [10]	10.21	0.634	12.43	0.551
Snowflake [5]	10.69	0.616	12.82	0.551
AdaPoinTr [6]	9.12	0.721	11.37	0.642
SVDFormer [28]	9.26	0.704	11.51	0.620
Ours	**8.71**	**0.723**	**10.49**	**0.665**

TABLE III
COMPARISON ON THE KITTI.

	SeedFormer [8]	AdaPoinTr [6]	3DMamba [12]	**Ours**
Fidelity↓	0.151	0.237	0.010	**0.008**
MMD↓	0.516	0.392	0.491	**0.377**

outperforms all comparative approaches in terms of CD-ℓ_1-Avg and F-Score@1% metrics across both seen and unseen categories, particularly exhibiting superior performance in the more challenging unseen category tests. This indicates that the noise discrimination and removal strategy based on latent feature space endows the model with enhanced robustness and generalization capability.

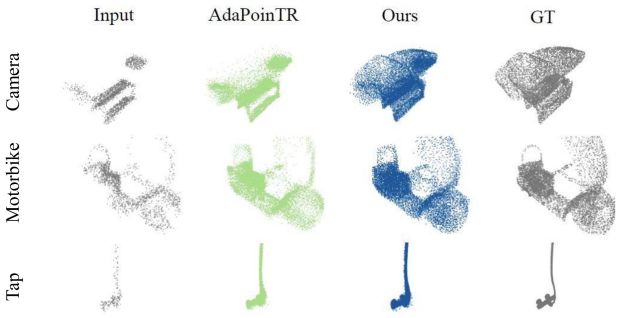

Fig. 4. Qualitative results on ProjectedShapeNet-55/34.

Results on KITTI. To evaluate the performance of the proposed method in real-world scenarios, we employ a pre-processing scheme to isolate automotive point clouds for independent testing. As demonstrated in Table III and Fig.5, our approach achieves superior results in both MMD and fidelity metrics, indicating closer alignment of the global distribution of generated point clouds with real-world data.

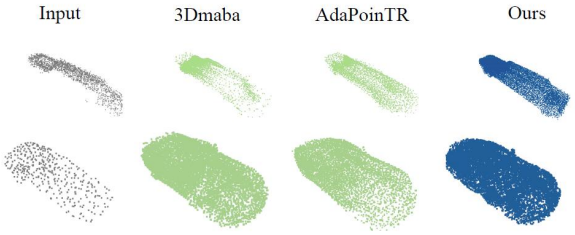

Fig. 5. More qualitative results on point cloud completion at KITTI. The pre-trained model weights are also made available in our open-source code.

TABLE IV
COMPLEXITY ANALYSIS. WE REPORT THEORETICAL FRAMES PER SECOND (FPS) AND BILLION FLOATING-POINT OPERATIONS (GFLOPS) OF OUR METHOD AND OTHER METHODS.

	GFLOPs	FPS
GRNet [3]	40.44	**39.13**
SnowflakeNet [5]	17.16	41.76
LAKeNet [21]	24.48	2.91
AdaPoinTr [6]	**12.43**	37.21
SeedFormer [8]	13.97	12.20
SPA-net(previous)	27.72	14.32
Ours	27.80	13.18

Complexity Analysis. In Table IV, we present a detailed complexity analysis of our method, comparing it with other methods in terms of Frames Per Second (FPS) and Billion Floating-point Operations (GFLOPs). While our method is not the most efficient in terms of computational complexity and speed, it strategically increases the computational cost to achieve the best performance on the benchmarks. The addition of the GAS and IGAR modules has a minimal impact on these two metrics compared to our previous method.

E. Ablation experiments

We perform ablation studies on Discriminator I, II, and the gradient affine transformation module. The experimental results are detailed in Table V.

In addition to the hypersphere hypothesis, as shown in Figure 6, the manifold hypothesis is also a common method for representing potential spaces. Therefore, in this section, the manifold hypothesis and the ratio of the distance from the boundary of the potential space to the normal boundary r_2/r_1

TABLE V
RESULTS ON PROJECTSHAPENET-55 VALIDATING THE EFFICIENCY ON KEY MODULES.

Discriminator I	Gradient Affine Transformation	Discriminator II	CD-ℓ_1-Avg↓	F-score@1%↑
✗	✗	✗	11.19	0.590
✓	✗	✗	10.81	0.613
✓	✓	✗	9.15	0.644
✓	✓	✓	**8.85**	**0.689**

were studied using the CD-ℓ_1-Avg($\times 10^3$) and F-score@1% metrics, with the results shown in Table VI. Compared to the two hypotheses, the CD-ℓ_1-Avg metric of the hypersphere hypothesis is generally better than that of the manifold hypothesis, but the F-score@1% is slightly better for the manifold hypothesis. This may be because the hypersphere hypothesis has a simpler structure, providing a clear constraint for normal data, making the representation of normal point clouds in the potential space more concentrated and compact. This helps to reduce reconstruction errors, thus performing better on the CD-ℓ_1-Avg metric.

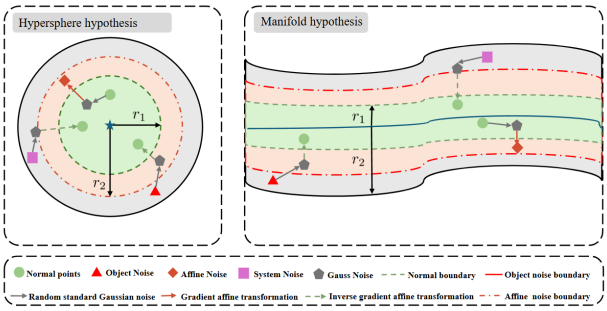

Fig. 6. Potential spatial representations of hypersphere hypothesis and manifold hypothesis. Hypersphere hypothesis. **Hypersphere hypothesis**: Assuming the data distribution is symmetrical and uniform, normal samples are strictly confined within the hypersphere, forming a convex envelope. **Manifold hypothesis**: Assuming high-dimensional data is distributed on a low dimensional manifold, the manifold locally approximates Euclidean space, but the overall structure may be nonlinear, non convex, or fragmented.

TABLE VI
THE HYPOTHESIS OF LATENT SPATIAL MODEL IS SET ON PROJECTSHAPENET-55 TO MELT EXPERIMENTAL RESULTS.

r_2/r_1	Hypothesis	CD-L_1-Avg	F-score@1%
1.5	Hypersphere	9.12	0.667
2	Hypersphere	**8.85**	0.689
3	Hypersphere	8.97	0.654
1.5	Manifold	9.37	0.686
2	Manifold	9.19	0.695
3	Manifold	9.10	**0.701**

The boundary of the manifold is not as clear as that of the hypersphere. It may be more challenging to distinguish between normal and abnormal data, requiring more complex models to learn and recognize the boundaries of the manifold. However, this more complex boundary, under the premise of proper training, can better capture the complex structure and data distribution within it. The F-score@1% is often used to measure the model's precision at a low recall rate, reflecting the model's ability to accurately classify fine and edge points. Therefore, the manifold hypothesis is superior in recognizing fine details or edge points (F-score@1%). For the ratio of the distance from the affine boundary to the normal boundary r_2/r_1 , its effect on the model's performance is also significant. When this ratio is too small, the boundary may be smaller than the actual normal boundary, leading to an overly sensitive detector that easily misclassifies normal data as abnormal. Conversely, when this ratio is too large, the boundary becomes blurred, which may be beneficial for the approximation of the manifold boundary, as it reduces the restriction on the boundary range, providing more space for the complex structure of the manifold, but is not conducive to the approximation of the spherical boundary. The hypersphere hypothesis relies on the clear boundary to distinguish between normal and abnormal data, and a blurred boundary makes this distinction difficult, affecting the model's performance under the hypersphere hypothesis, especially in scenarios requiring precise boundary definition.

In summary, when using the hypersphere hypothesis and $r_2/r_1 = 2$, the best CD-ℓ_1-Avg($\times 10^3$)) result is 8.85, but the F-score@1% is a secondary result of 0.689. When using the manifold hypothesis and $r_2/r_1 = 3$, the best F-score@1% result is 0.701, but theCD-ℓ_1-Avg($\times 10^3$)) result is significantly lower than the former. Considering all factors, the hypersphere hypothesis,$r_2/r_1 = 2$ is chosen as the final model setting.

V. CONCLUSION

Given that point cloud completion tasks are highly sensitive to noise, we propose a novel noise synthesis and restoration strategy to alleviate this issue. Experimental results have demonstrated its effectiveness. We expect if to offer innovative insights for addressing noisy point cloud completion.

In the future, our aim is to explore methods to adapt SPA-Net to scene completion tasks, which would significantly expand its applicability. Additionally, we plan to explore methods to reduce the model's computational complexity, for instance, by investigating efficient attention mechanisms like FlashAttention, to achieve faster performance without sacrificing accuracy.

DATA AVAILABILITY

The data that support the findings of this study will be openly available at: *https://github.com/S2CTransNet/SPAn-net*

REFERENCES

[1] Xiaofei Qin, Lin Wang, Yongchao Zhu, Fan Mao, Xuedian Zhang, Changxiang He, and Qiulei Dong, "Rectified self-supervised monocular depth estimation loss for nighttime and dynamic scenes," *Engineering Applications of Artificial Intelligence*, vol. 144, pp. 110026, 2025.

[2] Suaib Al Mahmud, Abdurrahman Kamaruzzaman Rahiman, Azhar Mohd Ibrahim, and Ahmad Jazlan Haja Mohideen, "Advancements and challenges in mobile robot navigation: A comprehensive review of algorithms and potential for self-learning approaches," *Journal of Intelligent & Robotic Systems*, vol. 110, no. 3, pp. 120, 2024.

[3] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun, "Grnet: Gridding residual network for dense point cloud completion," in *Computer Vision – ECCV 2020,Lecture Notes in Computer Science*, Jan 2020, p. 365–381.

[4] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari, "Softpoolnet: Shape descriptor for point cloud completion and classification," in *Computer Vision – ECCV 2020,Lecture Notes in Computer Science*, Jan 2020, p. 70–85.

[5] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han, "Snowflake point deconvolution for point cloud completion and generation with skip-transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6320–6338, Apr. 2023.

[6] Xumin Yu, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou, "Adapointr: Diverse point cloud completion with adaptive geometry-aware transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14114–14130, 2023.

[7] Liang Pan, Xinyi Chen, Zhongang Cai, and .etc, "Variational relational point completion network for robust 3d classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11340–11351, 2023.

[8] Haoran Zhou, Yun Cao, Wenqing Chu, Junwei Zhu, Tong Lu, Ying Tai, and Chengjie Wang, "Seedformer: Patch seeds based point cloud completion with upsample transformer," in *Computer Vision – ECCV 2022*, Cham, 2022, pp. 416–432, Springer Nature Switzerland.

[9] Xiaofei Qin, Anluo Yi, Wei Wang, Changxiang He, Lin Wang, Shiwei Tao, Xuedian Zhang, and Qiulei Dong, "SPA-Net: Skeletal-to-whole point cloud completion via progressive off-attention neighboring feature aggregation," *IEEE Transactions on Multimedia*, 2026.

[10] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou, "Pointr: Diverse point cloud completion with geometry-aware transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12478–12487.

[11] Liang Pan, Tong Wu, and Zhongang Cai .etc, "Multi-view partial (mvp) point cloud challenge 2021 on completion and registration: Methods and results," 2021.

[12] Yixuan Li, Weidong Yang, and Ben Fei, "3dmambacomplete: Exploring structured state space model for point cloud completion," 2024.

[13] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, vol. 27.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, vol. 33.

[16] Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang, "A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization," in *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, Eds., Cham, 2025, pp. 37–54, Springer Nature Switzerland.

[17] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu, "FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows," *arXiv e-prints*, p. arXiv:2111.07677, Nov. 2021.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2023.

[19] Yijun Long, Zhaoyu Chen, Hong Lu, and Wenqiang Zhang, "Gsformer: Geometric-spatial transformer on point cloud completion," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 1175–1180.

[20] Liang Pan, Xinyi Chen, Zhongang Cai, Junzhe Zhang, Haiyu Zhao, Shuai Yi, and Ziwei Liu, "Variational relational point completion network," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8524–8533.

[21] Junshu Tang, Zhijun Gong, Ran Yi, Yuan Xie, and Lizhuang Ma, "Lake-net: Topology-aware point cloud completion by localizing aligned keypoints," 2022.

[22] Binfu Ge, Shengyi Chen, Weibing He, Xiaoyong Qiang, Jingmei Li, Geer Teng, and Fang Huang, "Tree completion net: A novel vegetation point clouds completion model based on deep learning," *Remote Sensing*, vol. 16, no. 20, 2024.

[23] Shanshan Li, Pan Gao, Xiaoyang Tan, and Mingqiang Wei, "Proxyformer: Proxy alignment assisted point cloud completion with missing part sensitive transformer," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 9466–9475.

[24] Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, and Guanyu Yang, "Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 6070–6079.

[25] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, 2019, KDD '19, p. 2828–2837, Association for Computing Machinery.

[26] Guo Pu, Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian, "Controllable image synthesis with attribute-decomposed gan," in *Pattern Analysis and Machine Intelligence (TPAMI), 2022 IEEE Transactions on*, 2022.

[27] Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang, "A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization," in *European Conference on Computer Vision*. Springer, 2024, pp. 37–54.

[28] Zhe Zhu, Honghua Chen, Xing He, Weiming Wang, Jing Qin, and Mingqiang Wei, "Svdformer: Complementing point cloud via self-view augmentation and self-structure dual-generator," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 14508–14518.

[29] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "KITTI vision benchmark suite," 2012.

Architectures of previous work

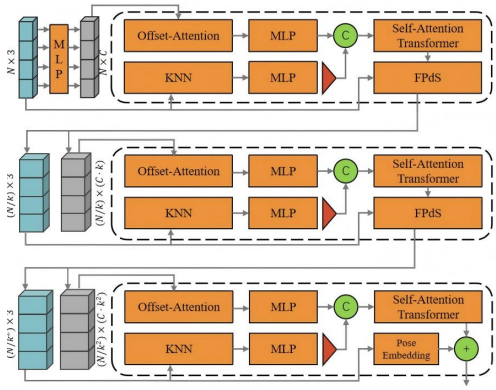

Fig. 7. Architecture of MSNFA. It provides three scales of features through three stages of feature extraction and two feature aggregation processes.

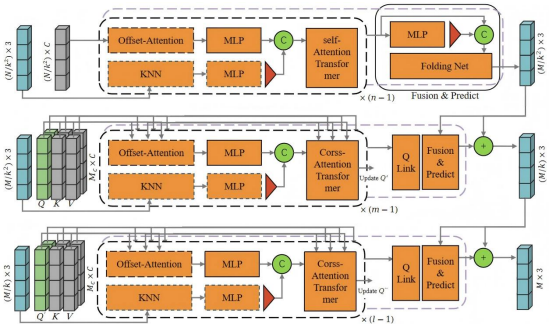

Fig. 8. Architecture of the skeletal-to-whole rebuilding module, consisting of three rebuilders. The first rebuilder utilizes the encoder's final layer output, while the subsequent rebuilders incorporate intermediate or final decoder layers along with a Q-Link. Each stage progressively refines the point cloud by applying predicted offsets to the preceding level's output.