A POINT CLOUD COMPLETION NETWORK VIA THE LATENT SPACE-DRIVEN TWO-STAGE NOISE SYNTHESIS AND RESTORATION STRATEGY

Xiaofei Qin, Anluo Yi, Jie Zhang, Shiwei Tao

University of Shanghai for Science and Technology
Shanghai, 200093, China
xiaofei.qin@usst.edu.cn
{232260517, 233370860, 243370925}@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

1. INTRODUCTION

Point cloud completion aims to rectify occlusions, sampling constraints, or sparsity-induced missing points, providing high-quality inputs for downstream tasks [1, 2, 3]. Its core lies in learning implicit 3D distributions to transition from partial observations to complete geometric models [4, 5, 6, 7]. 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 [4, 5, 8] and those based on Transformer methods[6, 7, 9].

This work was supported by the National Key R&D Program of China under Project [2021YFB2802300]

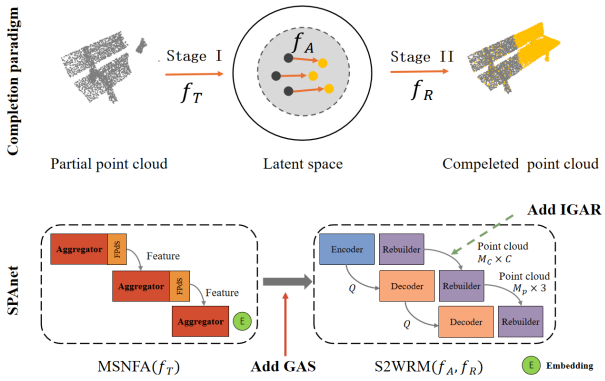

Fig. 1. Architecture of two-stage point cloud completion paradigm and SPA-net. **Paradigm** map known points $P_{know}$to latent space $Z_{konw} \subseteq \mathbb{R}^d$. In reconstruction, they get missing point cloud features $f_{\mathcal{A}}(Z_{\text{know}})$ from known ones in latent space and decode 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$\{\mathcal{P}_1 \in \mathbb{R}^{M_c \times 3}, \mathcal{P}_2 \in \mathbb{R}^{M_p \times 3}, \mathcal{P}_3 \in \mathbb{R}^{M \times 3}\}, M : M_p : M_c = 16 : 4 : 1$.

Recently, Transformer-based point completion methods have gained significant attention for 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 (detailed in **Appendix**), we proposed SPA-Net (second line of Fig.1) achieved state-of-the-art performance on ShapeNet-55/34[10] and MVP[11] benchmarks. However, our previous work, as well as most existing Transformer-based representative baselines such as

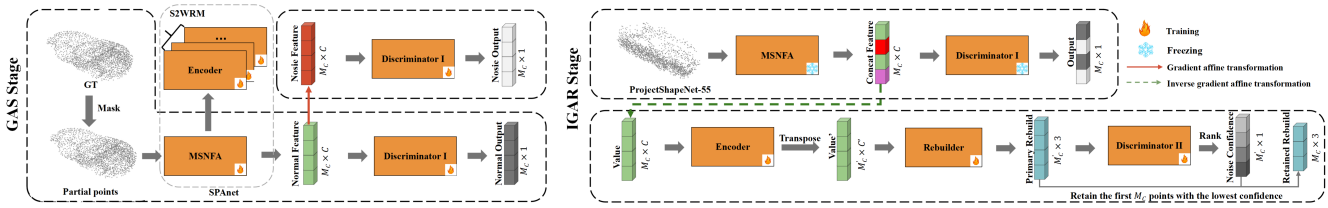

Fig. 2. Architecture of SPAn-Net. SPAn-Net additional incorporates two sequentially trained stages: The **GAS stage** then simulates missing data by randomly masking the ground truth point cloud from ProjectShapeNet-55, using MSNFA to process these partial points. It employs gradient affine transformations to create pseudo-noise points for training Discriminator I, which learns noise discrimination. In the **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.

PionTr[10], 3DMamba[12], and SeedFormer[9]—lack specific optimization for point cloud noise in practical scenarios, rendering them noise-sensitive and unable to fully meet deployment/downstream task demands. Additionally, the scarcity of realistic noisy public datasets further limits related research and applications.

In generative models [13, 14, 15], the latent space plays a pivotal role in noise suppression and anomaly detection. Per 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).

2. METHOD

2.1. Model architecture

To mitigate these challenges, GAS first tackles data scarcity by generating pseudo-noise via gradient affine transformations on normal embeddings, synthesizing diverse training scenarios that enable the model to learn recognizing and handling various noise types without extensive real-world data. Complementarily, IGAR performs soft correction on noisy point embeddings, preserving valuable geometric cues instead of discarding potentially useful 3D coordinates to restore synthesized noisy points to their correct positions, ensuring the final point cloud is both accurate and consistent with the expected data distribution.

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 ($\hat{r} \to r_1$) 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 $\mathcal{G} \in \mathbb{R}^{8192\times3}$ 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 ($\mathcal{D}_1(u_i) \to 0$). Conversely, for each affine noise point embedding $v_i \in U_v$ (generated via gradient-based transformations introduced in 2.2), 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. **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:

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

$x_i \in \mathcal{X}_n$ 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 2.2. 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 $U'_c \in \mathbb{R}^{M'_c\times3}$, then secondary discriminator $\mathcal{D}_2$ scores each point:

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

Finally, the top M_c points $\mathcal{P}_{refined} \in \mathbb{R}^{M_c\times3}$ with lowest noise scores are selected through ranking operations.

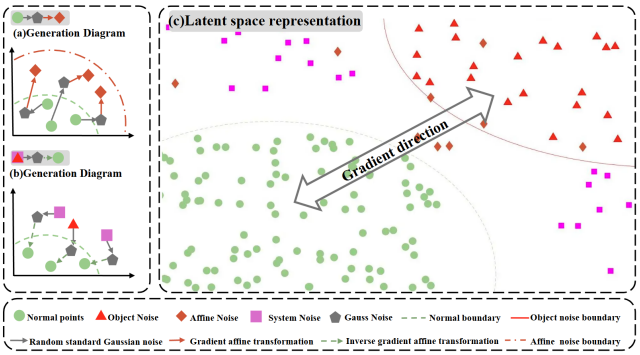

Fig. 3. Representation of latent space. (a) Gradient affine transformation. (b) Inverse gradient affine transformation. (c) Distribution of different point embeddedings

2.2. 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 limited to avoid the discriminator from 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: $u'_i = \tilde{u}_i - \widehat{\theta}_i$.

2.3. 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}:

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

3. EXPERIMENT

Our experiments span two noise-affected point cloud dataset types: simulated datasets with artificially injected noise

Table 1. Comparison on the ProjectedShapeNet-55. CD-ℓ_1-Avg denotes the average CD-$\ell_1 \times 10^3$ 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[4]	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[6]	10.49	6.35	11.20	12.59	15.24	10.07	8.12	7.49	11.34	0.594
AdaPoinTr[7]	8.81	5.18	9.77	10.89	13.27	8.81	6.79	5.58	9.58	**0.701**
Ours	**8.70**	**5.16**	**8.61**	**10.55**	**11.88**	**8.27**	**5.90**	**5.37**	**8.85**	0.689

Table 2. 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[4]	12.41	0.506	15.03	0.439
PoinTr[10]	10.21	0.634	12.43	0.551
Snowflake[6]	10.69	0.616	12.82	0.551
AdaPoinTr[7]	9.12	0.721	11.37	0.642
Ours	**8.71**	**0.723**	**10.49**	**0.665**

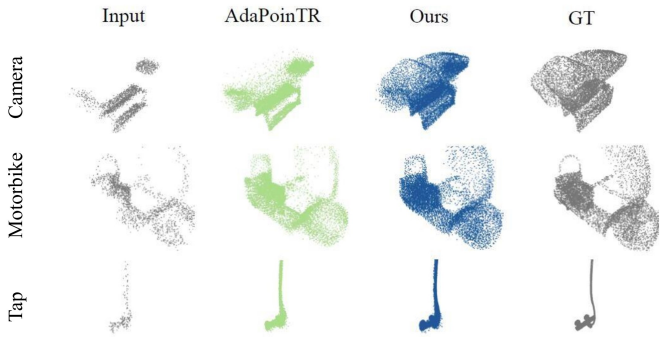

Fig. 4. Qualitative results on ProjectedShapeNet-55/34.

(ProjectedShapeNet-55/34[18], boasting rich data volumes and diverse categories highly recognized in the community), and the real-world KITTI dataset[19] with natural noise. Performance evaluation is carried out using widely adopted metrics. This approach enables a comprehensive assessment of our method's effectiveness and generalizability. Datasets, evaluation metrics, implementation details and more ablation studies are provided in the **Appendix**).

Results on ProjectedShapeNet-55. As shown in Table 1, 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 2 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 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.

Table 3. Comparison on the KITTI.

	SeedFormer[9]	AdaPoinTr[7]	3DMamba[12]	**Ours**
Fidelity↓	0.151	0.237	0.010	**0.008**
MMD↓	0.516	0.392	0.491	**0.377**

Table 4. 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**

Results on KITTI. To evaluate the performance of the proposed method in real-world scenarios, we employ a preprocessing scheme to isolate automotive point clouds for independent testing. As demonstrated in Table 3, 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.

Results of ablation study. We perform ablation studies on Discriminator I, II, and the gradient affine transformation module. The experimental results are detailed in Table 4.

4. 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.

5. REFERENCES

[1] Qiulei Dong, Zhengming Zhou, Xiaolan Qiu, and Liting Zhang, "A survey on self-supervised monocular depth estimation based on deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2025.

[2] 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.

[3] Suaib Al Mahmud, Abdurrahman Kamarulariffin, 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[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, "3dmamba-complete: 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] Liangliang Li, Guihua Liu, Feng Xu, and Lei Deng, "Carvingnet: Point cloud completion by stepwise refining multi-resolution features," *Pattern Recognition*, vol. 156, pp. 110780, 2024.

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

Appendix

Due to the page limit of the paper, we provide the appendix in the form of a link: *https://github.com/S2CTransNet/SPAn-net*