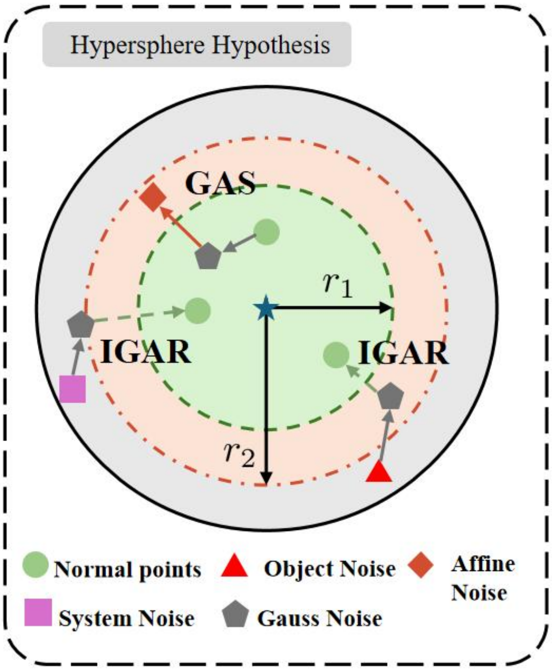
Submission Number: ICASSP-1347

Article Title: “A POINT CLOUD COMPLETION NETWORK VIA THE LATENT SPACE-DRIVEN TWO-STAGE NOISE SYNTHESIS AND RESTORATION STRATEGY”

Authors: Xiaofei Qin, Anluo Yi, Jie Zhang, Shiwei Tao

Reviewer#1, Concern # 1: The authors did not clearly explain the relationship and distinction between noise synthesis and noise restoration, nor why the method first synthesizes noise and then uses an inverse gradient affine transformation to restore it.

Author response: We appreciate your question regarding the relationship and distinction between noise synthesis and noise restoration. In our Transformer-based point-cloud completion pipeline, the two steps serve fundamentally different purposes and are tightly coupled to the latent-space representation learned by the multi-scale neighbor-feature aggregator (MSNFA) proposed in our prior work.



Why we first "synthesize" noise (GAS stage)

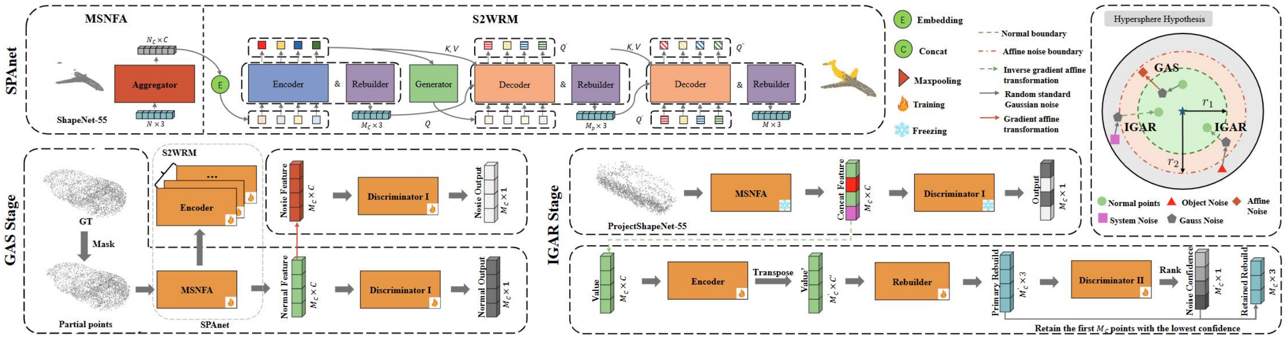
The MSNFA projects the input points into a high-dimensional latent semantic space in which all normal points of the same object lie approximately on a hypersphere (the "normal boundary"). To teach the first discriminator exactly where this boundary is, we need negative samples that are semantically close yet unambiguously outside the sphere. Instead of collecting extra real scans, we generate hard negatives on-the-fly by applying a gradient affine transformation to the normal embeddings. This operation moves the embeddings slightly beyond the normal boundary, yielding " pseudo-noise " that is indispensable for adversarial training. n short, synthesis is a data-augmentation step that makes the discriminator robust to both random and object-like noise.

Why we later "restore" with the inverse transformation (IGAR stage)

During inference, points that fall outside the learned boundary are regarded as noise. Simply deleting them removes possibly useful 3-D coordinates and breaks local surfaces. We therefore apply the inverse gradient affine transformation, mapping the out-of-sphere embeddings back into the normal region. The resulting "corrected" embeddings retain the original spatial cues but now comply with the normal point embeddings' distribution, so the decoder can safely attend to them without being perturbed. Thus, restoration is a soft correction that preserves information while guaranteeing statistical consistency with the latent-space model.

Reviewer#1, Concern # 2: Figure 2 is overly complex and fails to clearly illustrate the connections among the modules.

Author response: We sincerely apologize for the confusion caused by the complexity of Figure 2. However, due to the conference's restrictions on the length of the papers, we are unable to break down Figure 2 in more detail. The following is our detailed explanation of Figure 2.

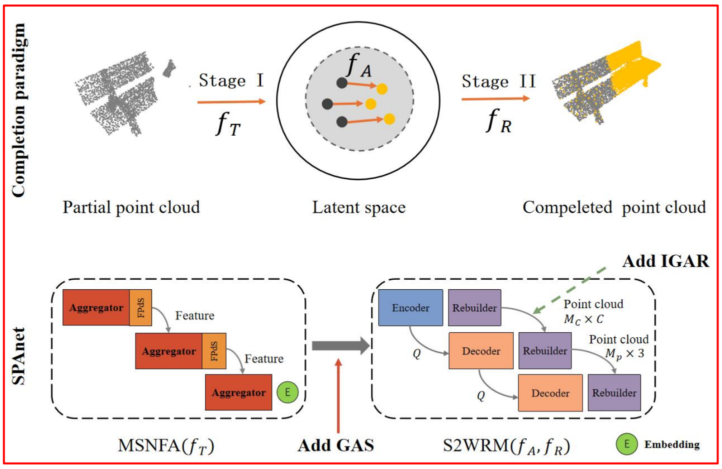


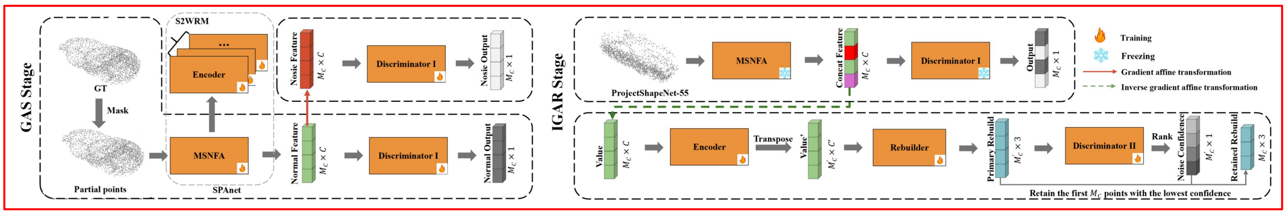
The diagram depicts an end-to-end pipeline that equips the original SPA-net we proposed with two extra, sequentially trained stages so it can reliably complete scans corrupted by noise. In this work, the SPA-net serves as the foundation for the model. It utilizes a Multi-Scale Neighbor Feature Aggregator (MSNFA) to extract features from the input point cloud.

Building upon the SPA-net, the GAS stage introduces a new branch to handle noisy point clouds. In this stage, the ground truth (GT) point cloud in Preojectshapenet-55, which is complete and noise-free, is randomly masked to simulate missing data. The MSNFA then extracts features from these partial points. The GAS stage applies a gradient affine transformation to synthesize pseudo-noise points, which are semantically similar to real noise. These noise points, along with the original clean points, are used to train Discriminator I, which learns to distinguish between noise and clean points.

In the IGAR stage of the pipeline, the model takes noisy, incomplete point clouds (train data of Preojectshapenet-55) as input and, while freezing the weights of the MSNFA and Discriminator I, generates an oversaturated point cloud that contains 125% of the expected points. This step is crucial for creating a dense point cloud that can be refined later. Discriminator II is then employed to filter out the noise, ensuring that only high-quality points that align with the expected number of points in a complete point cloud are retained. In summary, The Rebuilder reconstructs the point cloud, and the Ranking process evaluates the points based on their confidence scores. Points with the lowest confidence scores are recognized as noise and are subsequently removed, leaving behind only the most confident points, which are essential for the final, high-quality point cloud completion.

We split Figure 2 into two sub-figures in the hope of helping readers understand. The modified figures is as follows:





Reviewer#1, Concern # 3: In the Introduction, there is an "Anonymous statement," which is unusual since ICASSP is not a double-blind conference—this suggests the authors may have reused text from another submission without adequate revision.

Author response: We sincerely appreciate the reviewer for pointing out this critical issue, which has helped us further standardize the manuscript and avoid inappropriate content in the submission.

We acknowledge that we failed to carefully verify the review format of ICASSP (a non-double-blind conference) in the initial submission, and we mistakenly followed the double-blind review format to include the "Anonymous statement" in the Introduction. This was a careless oversight rather than an indication of reused text from other submissions.

We would like to clarify that the current work only adopts our previous research as the backbone network. The Gaussian Affine Synthesis (GAS) stage and Inverse Gradient Affine Restoration (IGAR) stage proposed in this paper are completely novel components, which achieve significant improvements in noisy point cloud completion compared with existing methods.

Regarding the materials related to our previous work and supplementary experiments, since we were informed by the editorial team that no appendix could be attached to the manuscript, we followed the editorial suggestions to provide a 9-page supplementary appendix and the source code via a designated link. The appendix includes a brief introduction to our previous work, detailed technical descriptions of the proposed GAS and IGAR stages, and comprehensive supplementary experimental results to support the claims in the main text.

We have already removed the inappropriate "Anonymous statement" from the Introduction and revised the relevant content to ensure full compliance with ICASSP's submission norms. We would like to emphasize again that **there is no situation of reusing text from other submissions without adequate revision in this manuscript.**

Reviewer#1, Concern # 4: The paper heavily depends on the authors' previous work, and several sections (including experiments) refer to an Appendix that is not actually provided, making it difficult to verify the claims.

Author response: We sincerely appreciate the reviewer's valuable comment. We wish to clarify that this work only adopts our previous research as the backbone network, while the GAS stage and IGAR stage are completely novel designs that drive the key improvements in noisy point cloud completion.

Regarding the missing appendix cited in the paper: upon consulting the editorial team, we were informed that no appendix could be attached directly to the manuscript. Following the editorial guidelines, we have provided a 9-page supplementary appendix and full source code via the following links for claim verification, and the links have also been added in the main text.:

Appendix link: ***https://github.com/S2CTransNet/SPAn-net***

Code link: ***https://github.com/S2CTransNet/SPA-net***

The appendix includes a concise overview of our prior work, detailed technical descriptions of the proposed modules, and comprehensive supplementary experimental results to support the findings in the main text. We confirm that this paper is self-contained with the supplementary materials, and does not rely on unprovided content to establish its conclusions.

Reviewer#2, Concern # 1: (1) clarification on the choice and tuning of key hyperparameters. (2) More details should be provided on training stability and the adversarial discriminators. (3) expansion on the experimental analysis of noise using different noise types and noise levels. There should also be explicit comparisons to noise-robust baselines

Author response: Thanks for your comments, we sincerely appreciate the reviewer for these insightful suggestions, which are crucial for improving the completeness and persuasiveness of our work. Due to the page limit of the conference, we have made some necessary revisions in the main text to facilitate readers' preliminary understanding, while the detailed content addressing all the above concerns has been included in our supplementary appendix.

For (1), detailed explanations are provided in Sections 3.3 and 3.5 of the appendix.

"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 $512 \times (1 + 0.25) = 640$.**"**

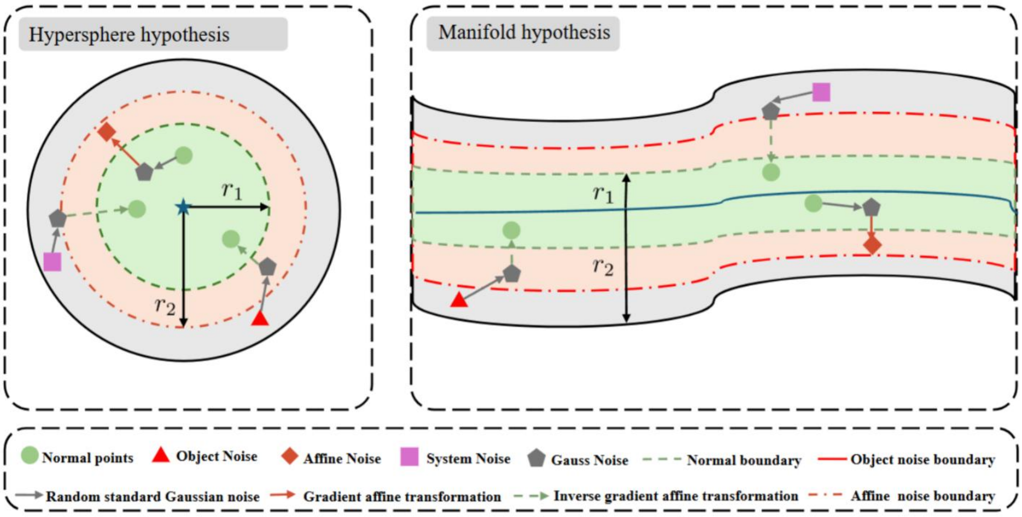


Fig. 8. Potential spatial representations of hypersphere hypothesis and manifold hypothesis. Hypersphere hypothesis. **Hypersphere hypothesis**: Assuming the data distribution is symmetrical and uniform, normal samples are strictly confined within the hypersphere, forming a convex envelope. **Manifold 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.

"In addition to the hypersphere hypothesis, as shown in Figure 8, 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 were studied using the CD-ℓ_1-Avg($\times 10^3$) and F-score@1% metrics, with the results shown in Table 3. 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 theF-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.

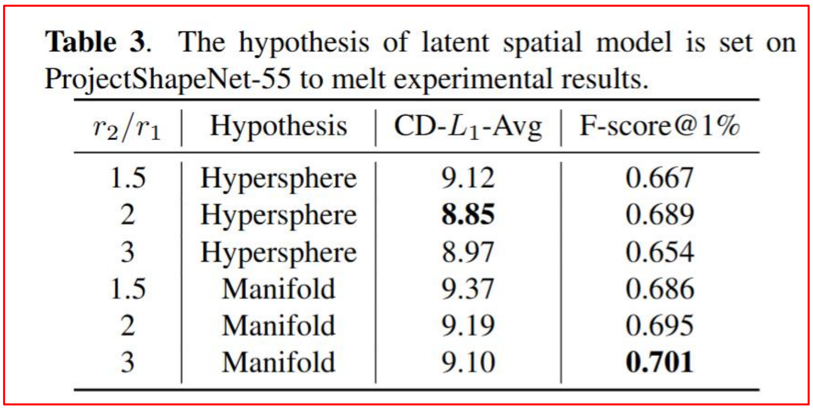


Table 3. The hypothesis of latent spatial model is set on ProjectShapeNet-55 to melt experimental results.

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 correctly classify fine or 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 the CD-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.”

For (2), detailed explanations are provided in Sections 2.3 of the appendix.

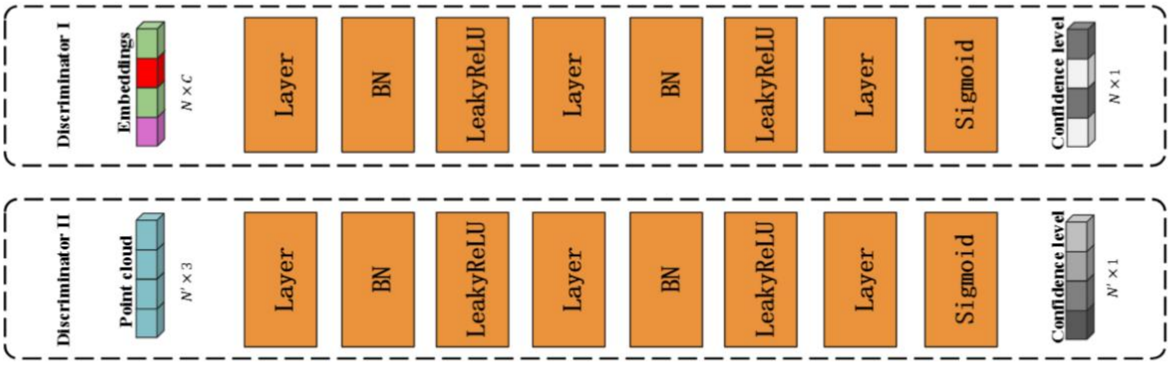


Fig. 6. Architectures of discriminators.

"As illustrated in Figure 6, the architectures $\mathcal{D}_1(\cdot)$ and $\mathcal{D}_2(\cdot)$ of both discriminators are identical in structure, differing only in the number of convolutional kernels within their fully connected layers."

For (3), We have added the quantitative comparison results of the Projectshapenet-34 dataset and the qualitative results of the KITTI dataset (Sections 3.4 in appendix). And the differences among these three datasets were elaborated in detail (Sections 3.1 in appendix). We believe that the types of data contained in these three datasets are rich enough.

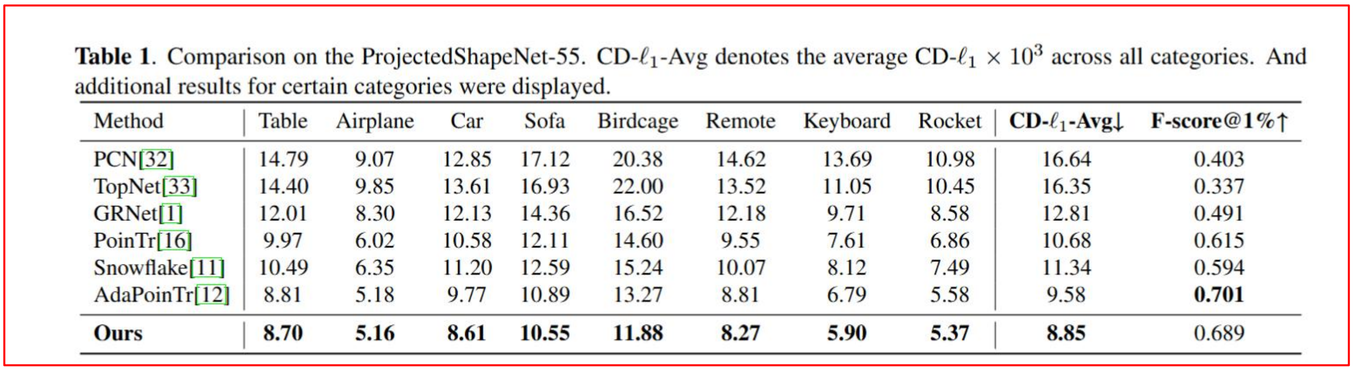


Table 1. Comparison on the ProjectedShapeNet-55. CD-ℓ_1-Avg denotes the average CD-ℓ_1 × 10^3 across all categories. And additional results for certain categories were displayed.

Method	Table	Airplane	Car	Sofa	Birdcage	Remote	Keyboard	Rocket	**CD-ℓ_1-Avg↓**	**F-score@1%↑**
PCN[32]	14.79	9.07	12.85	17.12	20.38	14.62	13.69	10.98	16.64	0.403
TopNet[33]	14.40	9.85	13.61	16.93	22.00	13.52	11.05	10.45	16.35	0.337
GRNet[1]	12.01	8.30	12.13	14.36	16.52	12.18	9.71	8.58	12.81	0.491
PoinTr[16]	9.97	6.02	10.58	12.11	14.60	9.55	7.61	6.86	10.68	0.615
Snowflake[11]	10.49	6.35	11.20	12.59	15.24	10.07	8.12	7.49	11.34	0.594
AdaPoinTr[12]	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

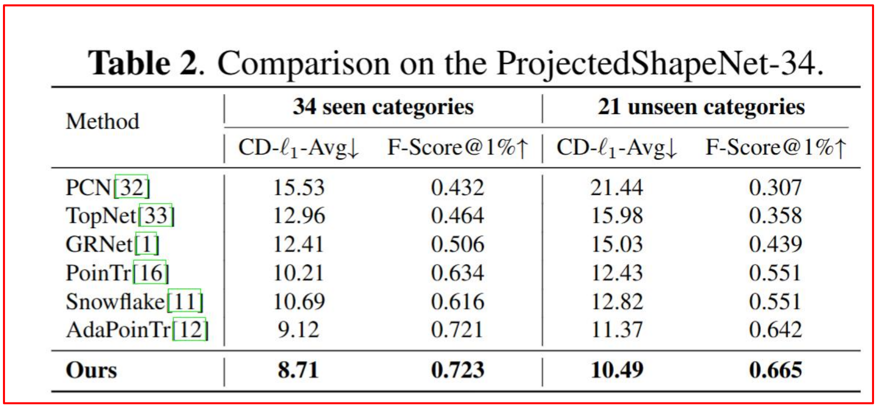


Table 2. Comparison on the ProjectedShapeNet-34.

Method	**34 seen categories**		**21 unseen categories**	
	CD-ℓ_1-Avg↓	F-Score@1%↑	CD-ℓ_1-Avg↓	F-Score@1%↑
PCN[32]	15.53	0.432	21.44	0.307
TopNet[33]	12.96	0.464	15.98	0.358
GRNet[1]	12.41	0.506	15.03	0.439
PoinTr[16]	10.21	0.634	12.43	0.551
Snowflake[11]	10.69	0.616	12.82	0.551
AdaPoinTr[12]	9.12	0.721	11.37	0.642
Ours	**8.71**	**0.723**	**10.49**	**0.665**

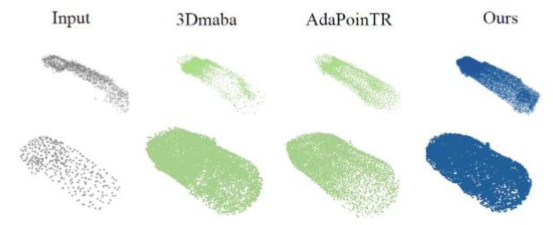


Fig. 7. More qualitative results on point cloud completion at KITTI. The pre-trained model weights are also made available in our open-source code.

"Table 1 and Table 2 present the comparison results between our method and other state-of-the-art approaches on the ProjectedShapeNet-55 and ProjectedShapeNet-34 datasets. It should be noted that the results for PoinTr \cite{4}, AdaPoinTr \cite{21}, and Snowflake \cite{20} are taken from their original papers. Since PCN \cite{11}, TopNet \cite{12}, and GRNet \cite{6} do not perform experiments on these datasets in their original papers, we use an existing paper \cite{21} to reproduce their methods and obtain results on the above datasets. Table 3 in manuscript shows the comparison results between our method and other state-of-the-art approaches on the KITTI dataset."

"ProjectShapeNet-55 and ProjectShapeNet-34 \cite{21} are alternative versions of the ShapeNet-55 and ShapeNet-34 benchmarks originally introduced by AdaPoinTr in TPAMI \cite{4}; the latter pair has become the most widely adopted and recognized evaluation suite in the point-cloud completion community.

To better mimic real-world acquisition, the authors additionally released ProjectShapeNet-55 and ProjectShapeNet-34. While ShapeNet-55/34 creates incomplete clouds by direct spatial cropping of complete ones, the “Project” variants are produced by noisy back-projection of single-view depth images: for every object, 16 random viewpoints are rendered, perturbed with depth noise, and back-projected to yield 16 distinct sparse and noisy partial clouds.

ProjectShapeNet-55 retains all object categories for both training and testing, whereas ProjectShapeNet-34 employs only the 34 visible categories during training and evaluates generalization to the remaining, unseen categories. All four datasets share identical ground-truth complete point clouds, enabling fair comparison.

KITTI \cite{69} 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.**”**

Reviewer#3, Concern # 1: The paper proposes a solution for point cloud completion. The novelty of the scheme relates to noise synthesis followed by a restoration strategy. The experimental results are encouraging. The paper is difficult to comprehend with many details of the proposed scheme not explained in sufficient detail.

Author response: We sincerely appreciate your insightful comments and valuable feedback on our manuscript. We have noticed that your concerns are partially overlapping with those raised by another reviewer, which helps us pinpoint the key aspects needing improvement.

We acknowledge that the descriptions of the method details in the main text are not sufficiently comprehensive, which is mainly constrained by the page limit of the conference. To address this issue and enhance the readability of the manuscript, we have made targeted revisions to the main text, including simplifying the presentation of core frameworks and adding concise annotations for key modules.

For the detailed elaboration of the proposed scheme, including the design rationale of the noise synthesis and restoration strategy as well as the technical details of each module, we have provided a complete and thorough description in our supplementary appendix. We hope these supplementary materials can help you gain a more comprehensive and in-depth understanding of our work.
