Title:
Adaptive Point Cloud Denoising via Attention-Based Hard Masking and Variable-Length Feature Fusion

Authors:
Wen-Peng Ge, gewenpeng22@mails.ucas.ac.cn, University of Chinese Academy of Sciences
Li-Yong Shen, lyshen@ucas.ac.cn, University of Chinese Academy of Sciences

Keywords:
Point Cloud Denoising, Attention Mechanism, Hard Masking, Variable-Length Feature Fusion, Adaptive Noise Filtering, Deep Learning

DOI: 10.14733/cadconfP.2025.261-266

Introduction:
Point clouds, as a 3D data representation, are widely used in visualization, animation, CAD modeling, GIS, and autonomous driving due to their high spatial accuracy and detailed geometric characterization. Acquired via laser scanning, depth cameras, or stereo vision, point clouds inherently suffer from noise caused by environmental interference, sensor limitations, and measurement errors. Such noise degrades data quality and negatively impacts downstream tasks like 3D reconstruction and robotic navigation, necessitating robust denoising techniques to enhance reliability.

Traditional denoising methods include bilateral filtering (preserving edges) [1], Moving Least Squares (MLS) (surface projection) [2], and Locally Optimal Projection (LOP) (geometry-consistent point redistribution) [3]. However, these methods often rely on manual parameter tuning, limiting adaptability. Recent advances focus on deep learning for its superior generalization. PCPNet-based architectures first eliminate outliers before denoising [4], while auxiliary tasks like normal filtering improve feature retention [5]. Reinforcement learning frameworks adaptively select denoising strategies using geometric priors [6]. Despite these advances, most learning-based methods still face two limitations: (1) Static feature aggregation: Existing attention mechanisms uniformly fuse features across all points, potentially propagating noise in high-uncertainty regions. (2) Fixed computational granularity: Methods often apply identical denoising operations to all patches, ignoring variations in local noise intensity and geometric complexity. In contrast, our work proposes a dynamic denoising framework that explicitly addresses these issues by introducing: (1) Hard feature masking to suppress noise-contaminated features in a spatially adaptive manner, and (2) Binary-gated fusion to enable variable-depth feature integration based on patch-specific conditions. This approach eliminates the need for manual parameter tuning while achieving parameter-free adaptation to both noise levels and geometric structures.

Main Idea:
Overview
Given a noisy point cloud $\hat{P}$, our method first predicts a noise displacement vector. The objective is to restore the point cloud $\hat{P}-\epsilon$ such that it closely approximates the clean point cloud P, where P accurately represents the true 3D geometry of the object while preserving sharp features at the edges. The restoration process can be formulated as:

$$\hat{P} = P + \epsilon .$$

We have designed two key modules, the Masked Sampling Module and the Variable-Length Feature Fusion Module, to enhance the predictive performance of our network. We present our network architecture in Fig. 1.

Network Architecture

Fig. 1 shows our complete network architecture. Given an input point cloud, for each point, a local patch is constructed by selecting its k-nearest neighbors. This local patch is then centralized and normalized to ensure consistency across different scales and locations. The input consists of a local patch formed by the N nearest neighbors centered at point $\hat{p}_i$, and the output is the noise displacement at point $\hat{p}_i$, which is used to correct the point p_i to lie closer to the true surface. Our network is mainly composed of four feature vector for processing by subsequent modules. The Masked Sampling Module adaptively masks out outliers with high noise. The Variable-Length Feature Fusion Module adaptively determines the number of feature fusion steps based on the information from the input patch. The Decoder reduces the high-dimensional features learned by the network into a 3D noise displacement vector, which is then used to correct the noisy point $\hat{p}_i$ to the true surface.

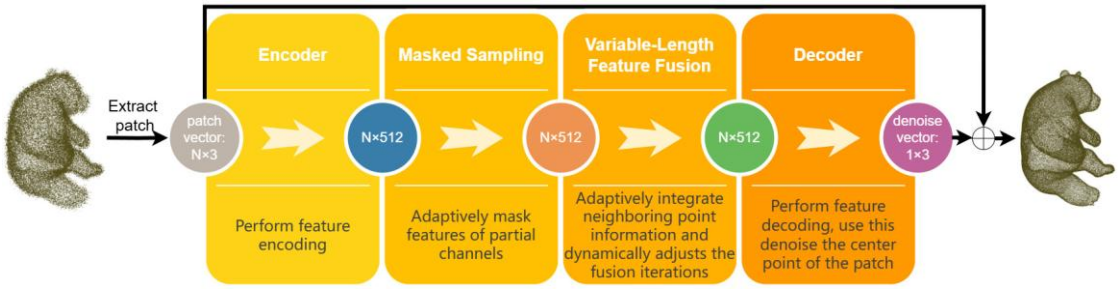


Fig. 1: Overview of network architecture.

The Encoder module begins by centering the input point cloud patch. It consists of four fully connected layers (MLPs) to get a high-dimensional feature representation of $N \times 512$, facilitating more efficient processing by subsequent deep learning modules. The Decoder module starts with a global max-pooling operation on the input feature tensor, generating a global descriptor that captures the patch's key structural features. This descriptor is then passed through three fully connected layers to produce a 3D noise displacement vector. The denoised coordinate is computed by subtracting the predicted displacement from the noisy coordinate p.

To mitigate the interference caused by high-noise outliers during feature fusion in point cloud denoising, we propose a Masked Sampling Module (MSM) to adaptively suppress outlier contamination. As illustrated in Fig. 2(b), the module operates as follows: 1. Feature Refinement with Residual MLP. Given the input feature embedding $f \in R^{N \times C}$, where N denotes the number of points and C represents feature dimensions, we first refine the features through a residual MLP sub-network. This architecture, designed to alleviate gradient vanishing issues, generates enhanced features $f_p \in R^{N \times C}$, which encapsulate both local geometric patterns and global contextual information. 2. Attention-Guided Feature Correlation Learning. To quantify inter-point feature dependencies within the patch, we introduce an attention-based scoring mechanism. Specifically, f_p is processed by an MLP layer followed by a sigmoid activation, yielding an adaptive significance score vector $f_s \in R^{N \times 1}$ with values in [0,1]. Higher scores indicate stronger relevance to the underlying surface structure, while

lower scores correspond to potential outliers. 3. Binary Mask Generation and Hard Masking. A threshold τ (empirically set to 0.5) is applied to binarize f_s, producing a binary mask $f_t \in \{0,1\}^N$. The mask is then expanded to match the feature dimensions of , enabling element-wise multiplication:

$$f_{masked} = f_p \otimes f_t .$$

This operation selectively nullifies features from outliers ($f_t = 0$), effectively decoupling their influence during subsequent feature aggregation stages. And the effectiveness of the hard masking technique is demonstrated in Fig. 2(a).

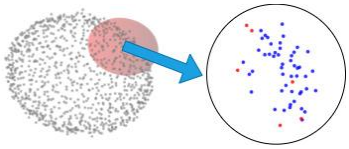


Fig. 2(a): Masking effect visualization.

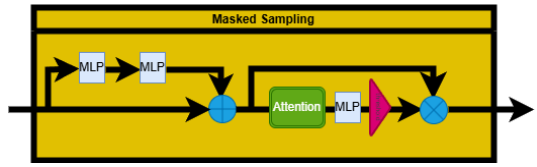


Fig. 2(b): Masked sampling module architecture visualization.

To adaptively aggregate multi-scale structural features while dynamically adjusting the depth of feature fusion based on local geometric complexity and noise distribution, we propose a Variable-Length Feature Fusion Module (VLFFM). As depicted in Fig. 3, the module operates through the following stages: 1. Feature Preprocessing via Residual MLP. The input feature $f \in R^{N \times C}$ is first processed by a residual MLP sub-network to enhance its adaptability to subsequent operations. This step ensures stable gradient propagation while refining both local geometric details and global contextual relationships. 2. Multi-Scale Feature Encoding. A max-pooling operation is applied to the refined features to generate a condensed global descriptor $f_g \in R^C$, which encapsulates the dominant structural characteristics of the patch. 3. Dynamic Fusion Depth Control. To adaptively determine the optimal fusion depth for varying geometric complexities and noise levels, f_g is passed through three cascaded MLP layers. The final layer employs a sigmoid activation to produce a 3-dimensional vector $s \in [0,1]^3$. A threshold τ (empirically set to 0.5) is applied to binarize s, yielding a ternary decision mask $m \in \{0,1\}^3$, where each binary value governs the activation of a corresponding feature fusion submodule. 4. Attention-Guided Feature Fusion. Each enabled submodule (indicated by $m_i = 1$) processes the input features through an MLP and max-pooling layer, followed by an attention mechanism that learns adaptive weights for feature aggregation.

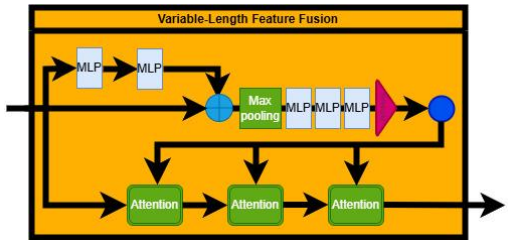


Fig. 3: Variable-length feature fusion module visualization.

Loss Function

The proposed loss function is formulated as a weighted combination of two complementary objectives to jointly optimize geometric fidelity and spatial distribution regularity. Specifically, the total loss L is defined as:

$$L = L_{recon} + \gamma L_{reg},$$

where L_{recon} enforces point-wise geometric consistency between the denoised coordinates and the ground-truth surface through an L2-norm penalty:

$$L_{recon} = \frac{1}{N}\sum_{i=1}^{N} \min_{j \in N(i)} \| \hat{p}_i - p_j^{gt} \|^2,$$

where $\hat{p}_i$ and p_i^{gt} denoting the denoised and ground-truth coordinates of the i-th point, respectively, $N(i)$ represents the k-nearest neighbors of point i. The regularization term L_{reg} imposes a uniformity constraint on the denoised point distribution to mitigate clustering artifacts and preserve surface continuity. This is implemented via a repulsion-based metric that penalizes abnormally dense regions:

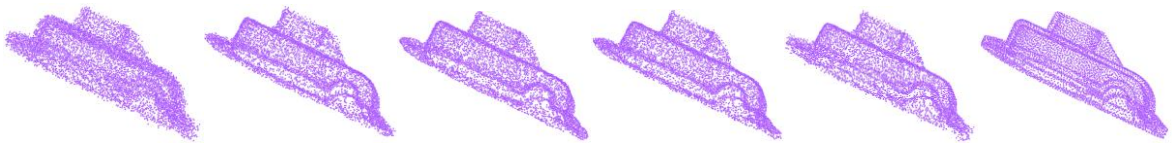
$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N} \max_{j \in N(i)} \| \hat{p}_i - p_j^{gt} \|^2.$$

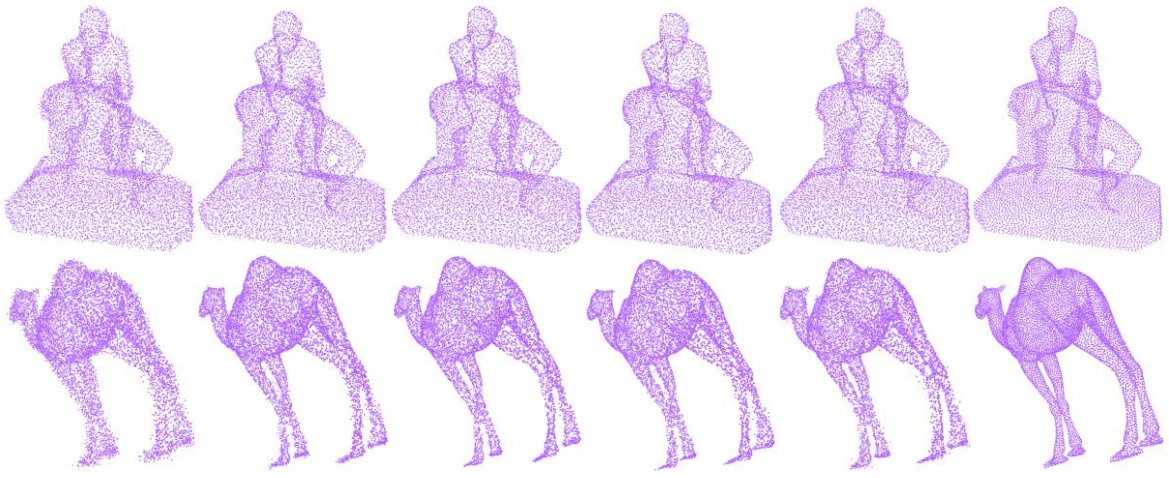
Results

To rigorously evaluate our method's robustness, we conducted comprehensive comparative experiments against three state-of-the-art denoising approaches: POINTCLEANNET [4], PCDNF [5], and PathNet [6]. Benchmarking was performed on the synthetic Synth-A dataset originally introduced in PathNet [6]. As quantitatively summarized in Tab. 1, our framework establishes new performance records across all evaluation metrics, achieving 3.6% lower Chamfer Distance (CD) than the second-best method. To validate generalization capabilities, we designed controlled experiments under varying conditions of noise intensity (δ = 0.5%–1.5% of bounding box diagonal) and point cloud density (10K–50K points per model). All evaluations adopted the CD metric. The visualization of our results is presented in Fig. 4, which demonstrates that our model excels in preserving finer details compared to other point cloud models. Specifically, the base of the car does not converge into a single layer, a feature that is not achieved by other existing models. Additionally, the gap between the legs of the humanoid statue is more accurately represented, closely resembling a realistic triangular shape. Furthermore, the details of the camel's ears and the denoising effect in the gap between its front legs are notably superior, highlighting the enhanced performance of our approach in capturing intricate structural details.

		Synth-A									
		10K			20K			50K			AVE
		0.5%	1%	1.5%	0.5%	1%	1.5%	0.5%	1%	1.5%	
CD	POINTCLEANNET	29.761	42.887	49.873	17.778	22.895	27.940	7.516	10.206	14.750	24.845
	PCDNF	27.167	39.779	46.904	16.176	21.638	26.319	7.072	9.677	**14.495**	23.247
	PathNet	27.997	41.171	48.033	16.955	21.933	27.055	7.155	9.876	14.500	23.853
	Ours	**25.098**	**37.865**	**45.183**	**15.267**	**20.949**	**25.693**	**6.858**	**9.499**	15.690	**22.456**

Tab. 1: Quantitative comparison on Synth-A.





(a) Input (b) POINTCLEANNET (c) PCDNF (c) PathNet (d) Ours (e) Truth

Fig. 4: Visualization of the results.

Conclusions:

This work presents a robust point cloud denoising framework that addresses key limitations of existing methods through two novel contributions: the Masked Sampling Module for reduce the interference of outliers and the Variable-Length Feature Fusion Module fully leverage the latent structural information of neighboring points and it can adaptively determine the fusion depth based on the complexity of the point cloud. Extensive experiments on synthetic datasets demonstrate state-of-the-art performance, with our method outperforming POINTCLEANNET, PCDNF, and PathNet by 3.6% in Chamfer Distance under varying noise levels (0.5%–1.5%) and point densities (10K–50K points). The proposed dynamic depth control mechanism proves particularly effective in preserving fine structures under sparse sampling conditions, as validated by visual and quantitative analyses.

Acknowledgements:

This work was partially supported by Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDB0640200; the National Natural Science Foundation of China under Grant 12371384 and Fundamental Research Funds for the Central Universities.

Wen-Peng Ge, https://orcid.org/0009-0002-2536-6060
Li-Yong Shen, https://orcid.org/0000-0001-5769-4814

References:

[1] Digne, J.; De Franchis, C.: The bilateral filter for point clouds, Image Processing On Line, 7, 2017, 278-287. https://doi.org/10.5201/ipol.2017.179
[2] Alexa, M.; Behr, J.; Cohen-Or, D.: Computing and rendering point set surfaces, IEEE Transactions on visualization and computer graphics, 9(1), 2003, 3-15. https://doi.org/10.1109/TVCG.2003.1175093
[3] Lipman, Y.; Cohen-Or, D.; Levin, D.: Parameterization-free projection for geometry reconstruction, ACM Transactions on Graphics (ToG), 26(3), 2007, 22-es. https://doi.org/10.1145/1276377.1276405
[4] Rakotosaona, M. J.; La Barbera, V.; Guerrero, P.: Pointcleannet: Learning to denoise and remove outliers from dense point clouds, Computer graphics forum, 39(1), 2020, 185-203. https://doi.org/10.1111/cgf.13753

[5] Liu Z.; Zhao Y.; Zhan S.: PCDNF: Revisiting learning-based point cloud denoising via joint normal filtering, IEEE Transactions on Visualization and Computer Graphics, 2023. https://doi.org/10.1109/TVCG.2023.3292464

[6] Wei Z.; Chen H.; Nan L.: PathNet: Path-selective point cloud denoising, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. https://doi.org/10.1109/TPAMI.2024.3355988