2025 IEEE International Conference on Cloud Computing Technology and Science (CloudCom) | 979-8-3315-6634-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/CLOUDCOM67567.2025.11331312

Sparse Manifold Retrieval Network for ICESat-2 Photon Point Cloud Denoising

Hengming Dai[1,2,3], Haihan Duan[4,5], Cong Zhang[4,5], Yang Zhou[6], Xiaoyi Fan[4,5,*], Zhifang Zhao[1,2,3,*]
[1]Institute of International Rivers and Eco-Security, Yunnan University, Kunming 650500, China
[2]School of Earth Sciences, Yunnan University, Kunming 650500, China
[3]Yunnan International Joint Laboratory of China-Laos-Bangladesh-Myanmar Natural Resources Remote Sensing Monitoring, Kunming, 650500, China
[4]Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Guangdong, China
[5]Guangdong-Hong Kong-Macao Joint Laboratory for Emotion Intelligence and Pervasive Computing, Guangdong, China
[6]Nova Stella (Shenzhen) Technology Co., Ltd., China
Email:{daihm@ynu.edu.cn, duanhaihan@smbu.edu.cn, zhangcong@gml.ac.cn zhouyang@novastella.com.cn, xiaoyi.fan@smbu.edu.cn, zhaozhifang@ynu.edu.cn}

***Abstract*—The photon point clouds acquired by ICESat-2/ATLAS offer unprecedented potential for Earth observation but are heavily contaminated by noise photons, posing a significant challenge for downstream applications. Traditional denoising methods, which often rely on local density statistics, struggle with complex terrains and varying signal-to-noise ratios. While deep learning presents a promising alternative, existing approaches often inefficiently process the inherently sparse data via 2D projections or non-optimized 3D networks. To address these limitations, this paper introduces a novel deep learning framework for ICESat-2 photon denoising, termed Sparse Manifold Retrieval Network (SMRNet). We propose a Manifold-Aware Convolution (MAC) module to capture the continuous manifold structures of signal photons through multi-scale dilated sparse convolutions, and a Cross-Scale Pyramid Enhancement (CSPE) module to effectively refine multi-level features extracted from the encoder. Evaluated on a manually annotated dataset covering southeastern coastal regions of China, SMRNet demonstrates superior performance over traditional denoising method and data-driven baselines across multiple metrics. The results underscore the effectiveness of SMRNet in enhancing denoising accuracy, particularly in challenging environments with sparse signals and rugged topography.**

***Index Terms*—ICESat-2, photon counting lidar, point cloud denoising, sparse convolution, deep learning**

I. Introduction

The Ice, Cloud, and Land Elevation Satellite-2 (ICESat-2), equipped with the Advanced Topographic Laser Altimeter System (ATLAS), represents a significant leap forward in Earth observation technology. It provides unprecedented high-precision elevation data for monitoring global ice sheets, estimating forest biomass, retrieving sea ice thickness, and mapping terrain, among other applications [1]. However, the exceptional detector sensitivity of ATLAS, while enabling single-photon level detection, also makes it highly susceptible to interference from solar background radiation and atmospheric scattering. As a result, the standard data product, ATL03 (Global Geolocated Photon Data), contains a substantial amount of random noise photons [2]. These noise photons are intermingled with signal photons reflected from the Earth's surface, significantly impairing the accuracy and reliability of downstream higher-level data products, such as the ATL08 land and vegetation elevation product. Consequently, the accurate extraction of valid signal photons from massive volumes of noise has become a crucial yet challenging preprocessing step in the ICESat-2 data processing chain.

Traditional denoising methods for ICESat-2 photon point clouds primarily rely on the prior knowledge that signal photons tend to cluster spatially while noise photons are distributed randomly. These methods can be broadly categorized into three types: clustering-based approaches (e.g., DBSCAN and its variants [3]), local statistics-based methods (e.g., local distance statistics [4]), and density threshold-based techniques (e.g., the DRAGANN algorithm [5]). Although these conventional algorithms demonstrate acceptable performance under specific conditions—such as flat terrain and high signal-to-noise ratio (SNR)—their efficacy is highly dependent on predefined parameters (e.g., search radius, density threshold). Consequently, they often exhibit limited adaptability and weak generalization capability when applied to complex and variable terrain environments (e.g., dense forests, rugged mountains, coastal waters) or to data acquired under different illumination conditions (daytime vs. nighttime) [6, 7]. More specifically, in regions with inhomogeneous signal photon density distributions—such as within forest canopies or underwater topography—traditional methods are prone to erroneously removing low-density signal photons or misclassifying high-density noise near true signals as valid returns. These limitations can lead to significant biases in the extracted terrain and canopy height products [8].

In recent years, the rapid advancement of deep learning has provided new avenues to address this challenge. Its powerful capability for automatic feature extraction and complex pattern recognition has demonstrated significant advantages in processing high-dimensional and nonlinear data. Within the field of remote sensing, researchers have begun to explore the application of deep learning to denoise ICESat-2 data. Relevant approaches primarily employ 2D convolutional networks that take images as input, or point-based methods that

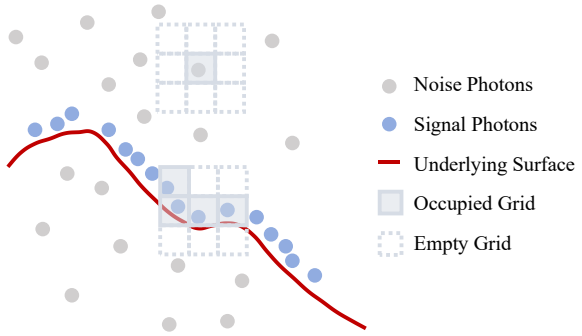

Fig. 1. Schematic Diagram of Photon Point Cloud. This figure illustrates a typical photon point cloud profile. The underlying signal photon trend is indicated by the red curve. A 3×3 sparse convolution kernel is also visualized within the diagram.

directly process 3D point clouds. These studies highlight the considerable potential and developmental prospects of deep learning-based denoising methods for ICESat-2 photon point clouds.

However, it is important to note that photon point clouds are inherently highly sparse and unordered sets of points, wherein noise photons are distributed randomly throughout the space, while signal photons exhibit local clustering characteristics. Moreover, at a global scale, signal photons collectively form a continuous manifold structure that aligns with the Earth surface and vegetation profile. Current data-driven denoising methods for photon point clouds have not yet fully explored this structural perspective.

Sparse convolution methods, on the other hand, can effectively exploit the sparsity of point clouds, reducing unnecessary computational overhead, while also offering advantages in representing local manifold structures. Therefore, this study introduces, for the first time, sparse convolution into the task of ICESat-2 photon cloud denoising. Building on this, we further propose a manifold-aware convolutional module that enhances the model's ability to recognize clustered photon signals through the integration of multi-scale sparse dilated convolutions.

The main contributions of this paper are as follows: (1) We pioneer the application of sparse convolution to the ICESat-2 denoising task and validate its effectiveness. (2) We propose a manifold-aware sparse convolution module and a cross-scale pyramid enhancement module to enhance the model's capability to capture the continuous distribution of signal photons. (3) We evaluate the proposed method on an ICESat-2 dataset acquired over the southeastern coastal regions of China, demonstrating its potential to surpass both traditional approaches and existing deep learning baselines in denoising performance.

II. Related Works

A. Photon Point Cloud Denoising

Denoising ICESat-2 photon point clouds constitutes a core challenge in data processing. Existing methods can be broadly categorized into traditional approaches and deep learning-based methods.

Traditional methods primarily rely on the prior that signal photons exhibit higher local density than noise photons, and perform filtering via threshold segmentation or neighborhood statistics [9]. For instance, Chen et al. [3] proposed an adaptive density discrimination elliptical filter, which uses along-track histogram statistics to separate surface and subsurface photons, thereby capturing more signals in deep-water regions. Leng et al. [10] employed kernel density estimation combined with a rotating elliptical neighborhood to improve land-water classification accuracy. Nan et al. [11] integrated local and global denossing through a two-stage filtering process to remove outliers and enhance accuracy. Yang et al. [4] introduced a backward elliptical distance (BED) mechanism that minimizes points within a rotating elliptical domain to better handle mountainous terrain with fluctuating signals. Zhang et al. [12] used a genetic algorithm to adaptively optimize denoising thresholds, mitigating the limitations of empirical parameter selection. Zhu et al. [13] applied the OPTICS algorithm [14], which is similar to DBSCAN but more robust to parameter choices. While these methods perform well in specific scenarios (e.g., underwater or mountainous regions), they often overlook the geometric and positional characteristics of signal photon neighborhoods. Moreover, they tend to struggle in environments with multi-modal density distributions, leading to either signal loss or noise misclassification, and generally lack adaptability to globally complex environments [15].

Deep learning-based approaches have recently been introduced for ICESat-2 denoising, leveraging data-driven strategies to capture complex patterns. Qin et al. [15] proposed converting each photon and its neighborhood into a 2D image and applied a GoogLeNet network integrated with a Convolutional Block Attention Module (CBAM) for classification, achieving higher accuracy than traditional methods in both simulated and real-data experiments. However, this method does not fully account for the inherent sparsity of point clouds. Liu et al. [9] directly processed 3D point clouds using VOJA-Net, which incorporates an ICESat Spatial Transformer (IS Transformer) and a Joint Attention Fusion (JA Fusion) module to enhance feature learning of the unique spatial distribution of ICESat-2 photons, yielding competitive performance. Meanwhile, Velikova et al. [7] explored the use of ConvPoint, a point-based neural network, trained directly on high-density lidar (HDL) data as ground truth labels. This approach effectively filtered outliers caused by low clouds in dense tropical rainforests and significantly improved the accuracy of terrain and canopy height estimates. Nevertheless, these methods either rely on fixed-size dense 2D grid conversion or employ network architectures that suffer from computational inefficiency when processing large-scale, highly sparse photon point clouds.

B. Deep Learning with Sparse Convolutions

With the widespread application of deep learning in 3D point cloud processing, the efficient extraction of both local and global features in sparse spaces has become a key research focus. Early approaches often converted point clouds into regular voxel grids and applied 3D convolutions directly for

feature learning [16]. However, due to the inherent sparsity of point clouds, such methods incur substantial computational redundancy from empty voxels and significant memory overhead. To address this issue, researchers have proposed various sparse convolution strategies, driving rapid advances in this area.

One early line of work is based on octree-structured hierarchical voxel representations, such as OctNet [17] and OCNN [18]. These methods partition the sparse point cloud space layer by layer via an octree and perform convolutions only on voxel nodes that contain points, thereby effectively reducing redundant computation. While these approaches achieved promising results in tasks such as 3D reconstruction and shape classification, they are limited by the complexity of maintaining the octree structure and suboptimal utilization of modern GPU parallel computing capabilities.

To further improve efficiency, Graham et al. [19] introduced Submanifold Sparse Convolution (SSC). Unlike standard sparse convolution, which generates new active points after convolution, SSC restricts computation to existing non-empty voxel locations and does not introduce new active units in the neighborhood. This design significantly mitigates the "expansion effect" of sparse tensors during convolution, enabling end-to-end training of deep networks on sparse point clouds while maintaining high efficiency. SSC has since become a cornerstone for subsequent research in sparse convolution.

Building on this, Choy et al. [20] proposed the Minkowski Engine, which provides a unified framework for high-dimensional sparse tensor computation. This framework supports both submanifold sparse convolution and regular sparse convolution, and can be flexibly extended to data of arbitrary dimensions (e.g., 2D, 3D, or 4D). Thanks to its efficient hash-based indexing and GPU-optimized implementation, the Minkowski Engine has been widely used in point cloud segmentation [21], object detection [22], and 3D scene understanding [23].

III. Methodology

A. Manifold-Aware Convolution Module

Although photon point clouds are inherently 3D data, they exhibit a strong strip-shaped distribution pattern. During manual annotation for denoising, these point clouds are often processed from a 2D perspective. Furthermore, 3D convolution over regular grids entails significantly higher computational costs compared to 2D convolution. Therefore, rather than directly processing the raw photon point cloud as three-dimensional data, we convert it into a 2D strip by defining the along-track distance (in meters) as the x-coordinate and the height above the WGS84 ellipsoid as the y-coordinate.

Under the strip-shaped 2D representation of photon point clouds, signal photons typically form continuous and approximately low-dimensional manifold structures within local regions. In contrast, noise photons tend to appear as isolated, sparsely scattered outliers. Consequently, an effective denoising strategy must not only capture fine-grained local spatial context, but also preserve potential manifold continuity across larger spatial scales.

However, traditional convolutional operations, constrained by a fixed receptive field, are inherently limited in modeling both local detail and global geometric patterns simultaneously. To overcome this limitation, we propose a Manifold-Aware Convolution (MAC) module that explicitly incorporates multi-scale context, as depicted in Fig. 2.

Formally, let the input and output 2D sparse tensors be denoted as $\mathcal{T}$ and $\mathcal{T}'$, they share the same spatial coordinates $\mathcal{C} \in \mathbb{Z}^2$, and $O_r = r * [-1, 0, 1]^2$ denotes the 3×3 convolution kernel offsets in 2D space, where r represents the corresponding dilation rate. For each non-empty location $(p, q) \in \mathcal{C}$, the 2D dilated convolution can be expressed as:

$$\mathcal{T}'[p, q] = \sum_{(i,j) \in O_r} \mathcal{W}[p - i, q - j] * \mathcal{T}[p - i, q - j] \quad (1)$$

where $\mathcal{W}$ is the convolution kernel weights. The MAC module employs k parallel dilated convolution branches, each with a distinct dilation rate $\{r_1, r_2, \ldots, r_k\}$. For the j-th branch, the feature tensor of the output $\mathcal{T}'$ is denoted as $\mathbf{F}_j \in \mathbb{R}^{N \times C_h}$, where N and C_h represent the number of non-empty positions in $\mathcal{T}'$ and the number of feature channels, respectively. The outputs from all dilated branches are concatenated along the channel dimension to form $\mathbf{F}_{\text{cat}} \in \mathbb{R}^{N \times (k \cdot C_h)}$, and subsequently generate a channel attention weight for cross-scale feature selection. At the same time, we also perform spatial aggregation over all $\mathbf{F}_j$ to obtain the spatial attention weights, where the use of multi-scale features enhances the reliability of spatial attention estimation.

By jointly applying channel attention and spatial attention across multiple dilation rates, the MAC module adaptively emphasizes informative responses while suppressing redundant or noisy activations. This design enables the network to effectively integrate local geometric structures with large-scale manifold continuity, leading to more robust denoising performance. Moreover, the multi-branch parallel design improves feature diversity without a significant increase in computational cost, achieving a favorable balance between accuracy and efficiency in photon point cloud denoising.

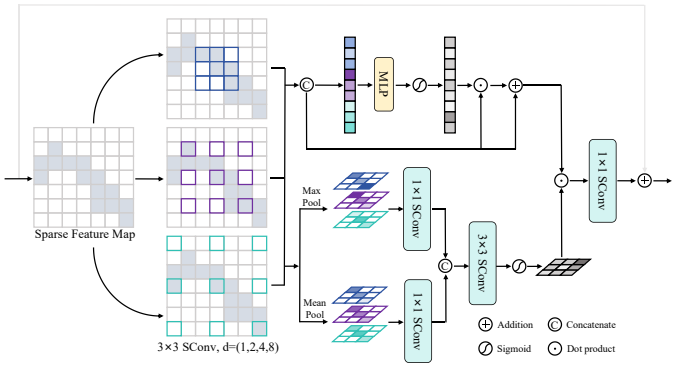

Fig. 2. Illustration of the proposed manifold-aware convolution module.

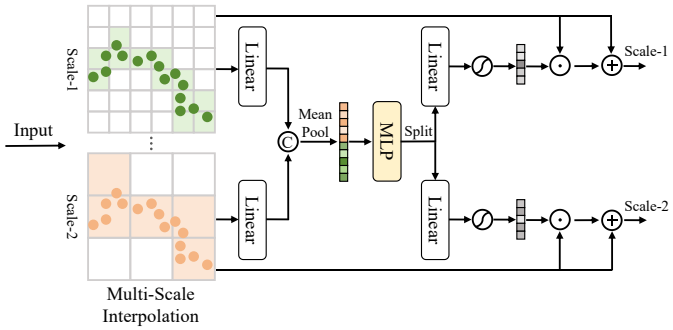

Fig. 3. Illustration of the proposed cross-scale pyramid enhancement module.

B. Cross-Scale Pyramid Enhancement

In photon cloud denoising, distinguishing signal photons from noise photons depends not only on information extracted from local neighborhoods but also on the integration of spatial context across multiple scales. Local neighborhoods provide fine-grained features that are particularly effective for identifying isolated background noise, since noise photons typically appear sparsely distributed without structural continuity. In contrast, large-scale spatial structures encode the global continuity of terrain surfaces, which is essential for discriminating true signal photons from noise in sparse or topographically complex regions. Therefore, an effective denoising framework must jointly exploit both local and global contextual information.

To explicitly integrate multi-scale contextual cues, a Cross-Scale Pyramid Enhancement (CSPE) module is introduced and seamlessly embedded into the U-Net backbone, implemented within the Minkowski Engine framework. Specifically, sparse tensor features obtained at different resolutions in the encoder are interpolated directly onto the original non-quantized photon point cloud to avoid the additional computational cost of learnable upsampling and to preserve precise geometric alignment with the raw spatial distribution. Subsequently, multi-scale features are concatenated to compute channel attention weights. This operation is conceptually similar to the multi-scale channel attention mechanism in the MAC module, as both aim to adaptively select informative features across scales. However, unlike MAC, the CSPE module first splits multi-scale features and generates channel attention weights for each resolution independently, addressing the challenge of directly fusing features with heterogeneous spatial densities. The enhanced features from each scale are then progressively fused in the decoder through hierarchical upsampling, ensuring coherent reconstruction of geometric structures across fine- and coarse-scale levels.

By complementing the Manifold-Aware Convolution (MAConv) module—which enhances intra-receptive-field feature extraction by capturing local manifold continuity—the CSPE module further reinforces cross-scale geometric awareness. This dual mechanism enables the network to effectively suppress discrete local noise while maintaining consistent recognition of continuous terrain surfaces at broader scales. Furthermore, by leveraging direct geometric interpolation instead of learnable upsampling, the CSPE design achieves a favorable balance between computational efficiency and denoising accuracy, demonstrating its suitability for large-scale photon point cloud processing.

C. Network Architectures

The proposed method adopts a U-Net-like architecture characterized by a symmetric encoder–decoder structure with skip connections to preserve fine-grained spatial information throughout the network, as illustrated in Fig. 4.

The encoder comprises four downsampling stages, each progressively reducing the spatial resolution of feature maps while expanding the receptive field. At each stage, two Manifold-Aware Convolution (MAConv) modules (as described in Section III-A) are stacked. This repeated application of MAConv enables the network to simultaneously capture local geometric structures and long-range contextual dependencies, thereby enhancing its ability to perceive continuous manifold patterns—such as terrain surfaces and vegetation layers—across extended spatial regions.

Before passing the multi-scale encoder features into the decoder for progressive upsampling, the Cross-Scale Pyramid Enhancement (CSPE) module (introduced in Section III-B) is employed to refine and balance feature representations from different depths of the encoder. Specifically, features at various resolutions are directly interpolated onto the original photon points, and cross-scale channel attention is applied to adaptively balance the model's focus across multiple scales. This design allows the decoder to reconstruct high-resolution representations that retain both semantic richness and spatial precision, significantly strengthening the model's capacity to integrate contextual information across scales.

By combining hierarchical feature extraction through MAConv in the encoder with multi-scale feature fusion via CSPE prior to decoding, the network effectively achieves comprehensive multi-scale information integration at both local and global levels. This joint design enhances the robustness of denoising while preserving the model's capability to accurately discriminate between signal and noise under complex and spatially heterogeneous conditions.

IV. Experiment

A. Dataset

The experiments in this study are conducted using the ATL03 global geolocated photon point cloud product from ICESat-2, with the study area located along the southeastern coast of China. The spatial extent of the study area is illustrated in Fig. 5. Since the signal confidence values provided in the ATL03 product cannot directly distinguish all potential signal photons, the acquired data were manually annotated using the PhotonLabeling[1] software [24]. Specifically, the

[1]https://github.com/zwshi-pku/PhotonLabeling

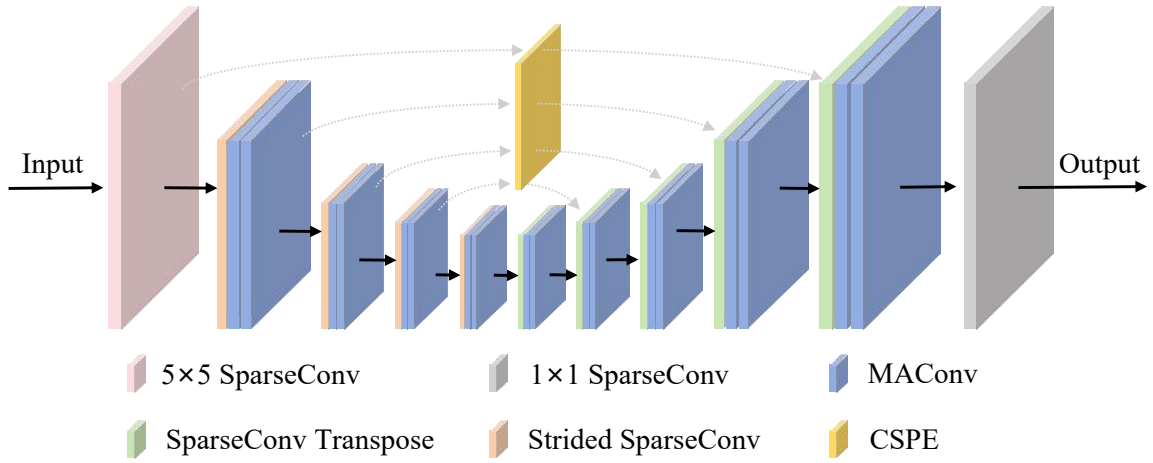

Fig. 4. Illustration of the overall network structure of the proposed SMRNet.

ATL03 photon point clouds were first projected into a two-dimensional strip space defined by along-track distance and height above the WGS84 ellipsoid (HAE). The spatial distribution of the annotated training and testing datasets is shown on the left side of Fig. 5, where red strips indicate training regions and blue strips represent testing regions. Each photon point cloud strip covers a length of 10 km, encompassing diverse terrain types including bare land, mountainous areas, and vegetated surfaces.

B. Evaluation Metrics

To comprehensively evaluate the performance of the proposed model in photon cloud denoising, several commonly used classification metrics were adopted for quantitative analysis, including Precision (P), Recall (R), F_1-score (F_1), Mean Intersection over Union ($mIoU$), Signal Photon IoU (IoU_1), Noise Photon IoU (IoU_2), and the Kappa coefficient (K). Let TP, TN, FP, and FN denote the number of correctly classified signal photons, correctly classified noise photons, signal photons incorrectly predicted as noise, and noise photons incorrectly predicted as signal, respectively. The corresponding evaluation metrics are defined as follows:

$$\begin{aligned} P &= \frac{TP}{TP+FP} & R &= \frac{TP}{TP+FN} \\ F_1 &= \frac{2\times P\times R}{P+R} & K &= \frac{P_o-P_e}{1-P_e} \\ IoU_1 &= \frac{TP}{TP+FP+FN} & IoU_2 &= \frac{TN}{TN+FP+FN} \end{aligned} \tag{2}$$

where P_o represents the observed agreement and P_e denotes the expected agreement by chance. For brevity, their detailed formulations are not presented here. These metrics collectively provide a comprehensive assessment of the model's denoising performance, reflecting its ability to accurately identify signal photons, suppress noise photons, and maintain overall classification consistency.

C. Implementation Details

To process long-span photon point cloud strips, a sliding window strategy was applied along the along-track direction with a window length of 1000 m. During model inference and training, different preprocessing strategies were adopted according to the network type. For sparse convolution-based methods (MinkowUNet and the proposed approach), all points within each segmented strip were quantized and directly fed into the model. The grid resolution was set to 5 m to balance spatial detail preservation and computational efficiency. For point-based method (PointNet++), 8192 points were randomly sampled from each segmented strip to enable batch-wise processing. In cases where the total number of photons was fewer than 8192, random sampling with replacement was applied to maintain a consistent number of input points.

All models were optimized using the AdamW optimizer with an initial learning rate of 0.1. A cosine annealing schedule was employed to gradually adjust the learning rate throughout the training process. Each model was trained for 50 epochs. All experiments were conducted on a workstation equipped with an Intel Core i9 CPU and an NVIDIA RTX 4090 GPU. The sparse convolutional networks were implemented using the PyTorch deep learning framework and the Minkowski Engine library, with CUDA version 11.8.

D. Experimental Results

To comprehensively assess the performance of the proposed method, we conducted comparative experiments against several representative baselines: the density-based clustering method DBSCAN (widely used for photon cloud denoising based on local point-density statistics), the point-based deep network PointNet++, and a sparse-convolutional semantic segmentation model, MinkowUNet. Qualitative visualization results are provided in Fig. 6, where each row shows the outputs of the different methods on the same test sample for direct comparison.

From the qualitative comparisons in Fig. 6, several observations can be made. The traditional density-based DBSCAN can

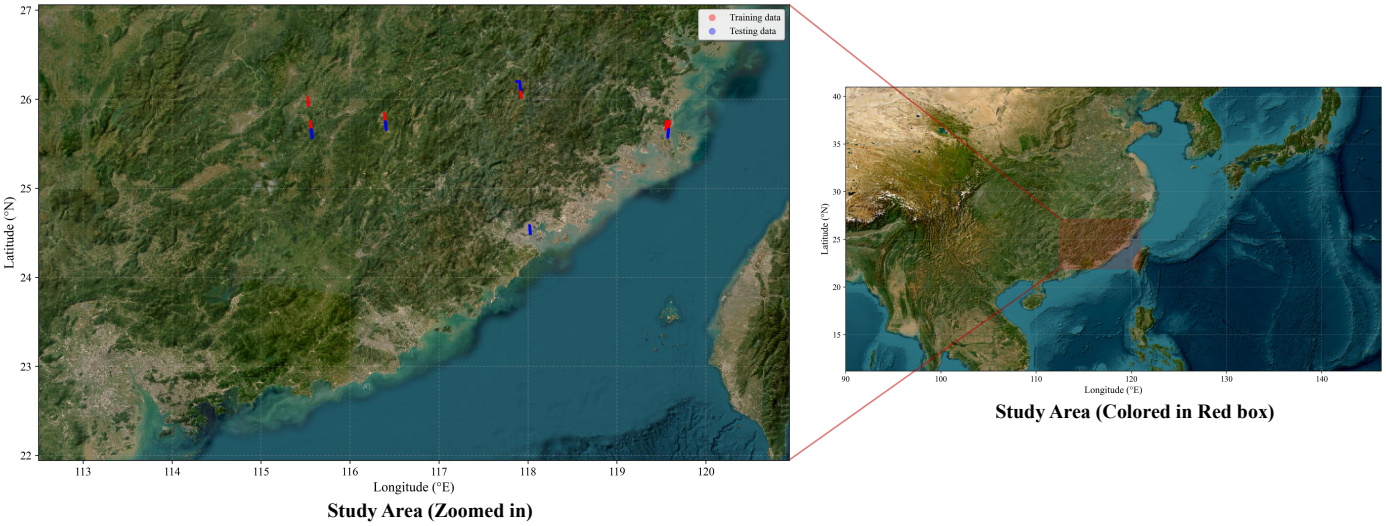

Fig. 5. Schematic diagram of the study area. The dashed line in the inset indicates the along-track scanning direction.

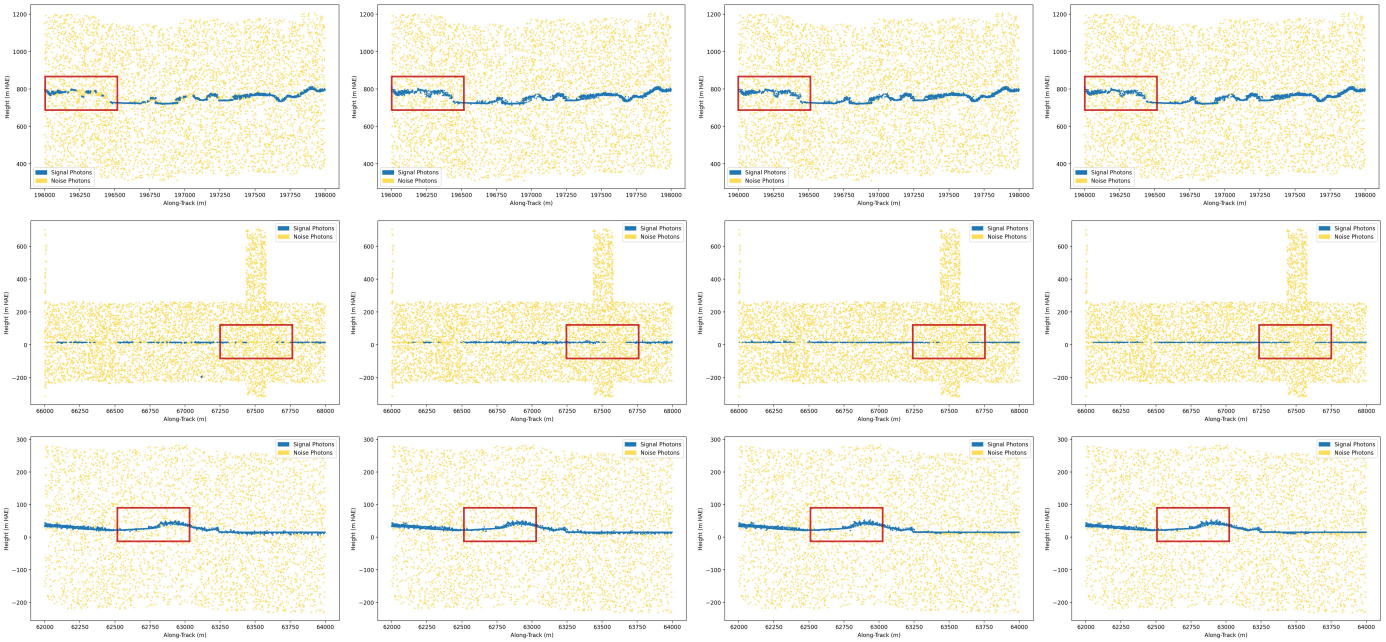

Fig. 6. Visualization results of photon point cloud denoising. The results of different methods are displayed in columns for each respective region, while each row shows the outcomes of various methods within the same region.

successfully identify most signal photons when there is a pronounced local density contrast between signal and background; moreover, it does not introduce substantial false alarms in such cases, indicating that DBSCAN is a reasonable choice when signal photons are relatively uniform and clearly denser than surrounding noise. However, DBSCAN's performance degrades markedly in more realistic scenarios where signal photon density varies. For example, in vegetated areas (Fig. 6, first row), canopy effects reduce point densities both at the ground and within the vegetation layer, causing DBSCAN to miss a significant portion of true signal photons. Similarly, in flat regions affected by atmospheric or cloud scattering (Fig. 6, second row), local signal continuity is disrupted and DBSCAN's reliance on simple density thresholds leads to reduced denoising accuracy.

Compared with DBSCAN, PointNet++ achieves noticeably improved denoising results—particularly in vegetation-covered regions—because learning-based methods can better extract discriminative local features and thus accommodate non-uniform signal distributions. Nevertheless, PointNet++ still exhibits limitations in modeling the continuity of signal photons over longer spatial extents (Fig. 6, second row) and

tends to produce locally coarse signal reconstructions. These shortcomings likely stem from the point-based processing paradigm: feature extraction and aggregation depend on local neighborhood searches, which makes PointNet++ more sensitive to variations in point density and less effective at capturing long-range contextual relationships.

The sparse-convolutional MinkowUNet demonstrates superior capability in preserving the continuity of signal photons relative to PointNet++ and DBSCAN, which validates the feasibility of employing 2D sparse convolution for photon cloud denoising. However, its outputs contain relatively more local residual noise compared with the proposed method. In contrast, the proposed method achieves a better balance between capturing continuous signal trends and suppressing local noise. This combined advantage indicates that our design more effectively integrates local feature extraction with multi-scale contextual fusion, enabling robust adaptation to varying signal densities and diverse terrain types.

Quantitative results on the test dataset are summarized in Table I. The proposed method attains leading scores across Precision, F_1, $mIoU$, IoU_1, IoU_2, and the Kappa coefficient, corroborating the qualitative observations and demonstrating superior overall denoising performance.

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT PHOTON POINT CLOUD DENOISING METHODS. BEST PERFORMANCES ARE MARKED IN BOLD.

Method	Metrics(%)						
	P	R	F_1	$mIoU$	IoU_1	IoU_2	K
DBSCAN	98.51	91.05	94.34	90.91	89.66	92.15	90.24
PointNet++	97.87	**97.96**	97.89	96.41	95.90	**96.92**	96.32
MinkowUnet	98.48	97.38	97.90	96.32	95.91	96.74	96.24
Ours	**98.56**	97.55	**98.03**	**96.52**	**96.15**	96.90	**96.45**

E. Module Effectiveness

In this section, we further validate the effectiveness of the MAC and CSPE modules through ablation experiments. The quantitative results are presented in Table II. To assess the contribution of each component, we constructed two ablation variants of the proposed network. In the first variant, the MAC module was removed and replaced with a standard sparse convolution layer with a kernel size of 3. In the second variant, the CSPE module was omitted, and the multi-scale features extracted from the encoder were directly forwarded to the decoder via skip connections, without cross-scale enhancement.

As shown in Table II, the removal of either the MAC or CSPE module results in a decline in overall performance, particularly in the F_1, $mIoU$ and K metrics. This confirms that both modules play essential and complementary roles-MAC effectively enhances local geometric feature extraction in irregular photon distributions, while CSPE strengthens multi-scale context fusion and structural consistency across layers. The observed performance degradation in their absence further substantiates the validity and robustness of the proposed architecture.

TABLE II
QUANTITATIVE EVALUATION OF MODULE EFFECTIVENESS. BEST PERFORMANCES ARE MARKED IN BOLD.

Model	Metrics(%)		
	F_1	$mIoU$	K
w/o MAC	97.89	96.34	96.26
w/o CSPE	97.83	96.19	96.09
Ours	**98.03**	**96.52**	**96.45**

V. CONCLUSION

This study presents a novel deep-learning-based framework to address the critical challenge of noise photon removal in ICESat-2 ATL03 data. The core strength of our approach lies in its ability to directly and efficiently process the inherently sparse structure of photon clouds. Specifically, the proposed Manifold-Aware Convolution (MAC) module expands the network's receptive field and enhances its capacity to distinguish the continuous, manifold-like distribution of signal photons from discrete noise. Meanwhile, the Cross-Scale Pyramid Enhancement (CSPE) module facilitates robust multi-scale feature fusion, enabling the model to jointly exploit fine-grained local details and broad contextual information.

We conducted extensive experiments using ATL03 photon point cloud data collected along the southeastern coast of China. The results demonstrate that the proposed method achieves superior performance compared with both traditional clustering-based approaches (e.g., DBSCAN) and representative data-driven baselines (e.g., PointNet++ and Minkowski-UNet). Importantly, this work provides the first empirical validation of the potential of sparse convolutional networks for photon point cloud denoising in ICESat-2 data.

However, due to the high cost of large-scale data screening and manual labeling, the dataset used in this study remains relatively limited in size and geographically localized compared with the global scope of ICESat-2 observations. In future research, we plan to expand the study area to explore typical large-scale application scenarios and establish a benchmark dataset for the ICESat-2 photon point cloud denoising task. Moreover, while the current approach achieves efficient denoising through a 2D strip-based representation and sparse convolutional processing, lightweight model design has not yet been fully considered. As part of our future work, we intend to introduce model compression and pruning strategies to develop a lightweight version of the framework, enabling deployment in cloud-based and edge-based environments for real-time or large-scale applications.

ACKNOWLEDGMENT

This work was supported in part by the Caiyun Postdoctoral Innovation Project in Yunnan Province under Grant C615300504106, and in part by the Key Field Projects of Ordinary Universities in Guangdong Province (No. 2025ZDZX3050).

REFERENCES

[1] T. Markus, T. Neumann, A. Martino, W. Abdalati, K. Brunt, B. Csatho, S. Farrell, H. Fricker, A. Gardner, D. Harding *et al.*, "The ice, cloud, and land elevation satellite-2 (icesat-2): science requirements, concept, and implementation," *Remote sensing of environment*, vol. 190, pp. 260–273, 2017.

[2] S. Popescu, T. Zhou, R. Nelson, A. Neuenschwander, R. Sheridan, L. Narine, and K. Walsh, "Photon counting lidar: An adaptive ground and canopy height retrieval algorithm for icesat-2 data," *Remote Sensing of Environment*, vol. 208, pp. 154–170, 2018.

[3] Y. Chen, Y. Le, D. Zhang, Y. Wang, Z. Qiu, and L. Wang, "A photon-counting lidar bathymetric method based on adaptive variable ellipse filtering," *Remote Sensing of Environment*, vol. 256, p. 112326, 2021.

[4] P. Yang, H. Fu, J. Zhu, Y. Li, and C. Wang, "An elliptical distance based photon point cloud filtering method in forest area," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[5] S. Gao, Y. Li, J. Zhu, H. Fu, and C. Zhou, "Retrieving forest canopy height from icesat-2 data by an improved dragann filtering method and canopy top photons classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[6] L. He, Y. Pang, Z. Zhang, X. Liang, and B. Chen, "Icesat-2 data classification and estimation of terrain height and canopy height," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103233, 2023.

[7] M. Velikova, J. Fernandez-Diaz, and C. Glennie, "Icesat-2 noise filtering using a point cloud neural network," *ISPRS Open Journal of Photogrammetry and Remote Sensing*, vol. 11, p. 100053, 2024.

[8] J. C. Fernandez-Diaz, M. Velikova, and C. L. Glennie, "Validation of icesat-2 atl08 terrain and canopy height retrievals in tropical mesoamerican forests," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2956–2970, 2022.

[9] Z. Liu, Y. Zi, X. Ma, Y. Ma, X. Wu, G. Jiang, H. Liu, F. Cheng, and Y. Zhou, "Voja-net: Vector-offset joint attention network for icesat-2 point cloud data denoising," *IEEE Geoscience and Remote Sensing Letters*, 2024.

[10] Z. Leng, J. Zhang, Y. Ma, J. Zhang, and H. Zhu, "A novel bathymetry signal photon extraction algorithm for photon-counting lidar based on adaptive elliptical neighborhood," *International Journal of Applied Earth Observation and Geoinformation*, vol. 115, p. 103080, 2022.

[11] Y. Nan, Z. Feng, B. Li, and E. Liu, "Multiscale fusion signal extraction for spaceborne photon-counting laser altimeter in complex and low signal-to-noise ratio scenarios," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[12] G. Zhang, W. Lian, S. Li, H. Cui, M. Jing, and Z. Chen, "A self-adaptive denoising algorithm based on genetic algorithm for photon-counting lidar data," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[13] X. Zhu, S. Nie, C. Wang, X. Xi, J. Wang, D. Li, and H. Zhou, "A noise removal algorithm based on optics for photon-counting lidar data," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 8, pp. 1471–1475, 2020.

[14] N. Febriana and I. Sitanggang, "Outlier detection on hotspot data in riau province using optics algorithm," in *IOP Conference Series: Earth and Environmental Science*, vol. 58, no. 1. IOP Publishing, 2017, p. 012004.

[15] W. Qin, Y. Song, Y. Zou, H. Zhu, and H. Guan, "A novel icesat-2 signal photon extraction method based on convolutional neural network," *Remote Sensing*, vol. 16, no. 1, p. 203, 2024.

[16] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. Ieee, 2015, pp. 922–928.

[17] G. Riegler, A. Osman Ulusoy, and A. Geiger, "Octnet: Learning deep 3d representations at high resolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3577–3586.

[18] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong, "O-cnn: Octree-based convolutional neural networks for 3d shape analysis," *ACM Transactions On Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.

[19] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9224–9232.

[20] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3075–3084.

[21] J. Li, H. Dai, H. Han, and Y. Ding, "Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 694–21 704.

[22] Z. Wang, Y.-L. Li, X. Chen, H. Zhao, and S. Wang, "Uni3detr: Unified 3d detection transformer," *Advances in Neural Information Processing Systems*, vol. 36, pp. 39 876–39 896, 2023.

[23] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 815–824.

[24] Z. Shi, J. Li, Z. Yang, H. Long, H. Cui, S. Zhao, X. Li, and Q. Li, "A linear feature-based method for signal photon extraction and bathymetric retrieval using icesat-2 data," *Remote Sensing*, vol. 17, no. 16, 2025.