A Novel Multidimensional Statistics Denoising Algorithm Based on Gaussian Mixture Model for Photon-Counting LiDAR Data

Sitong Chen, Sheng Nie, Xiaohuan Xi, Shaobo Xia, Feng Zhu, Cheng Wang, and Xiaoxiao Zhu

***Abstract*—The micropulse multibeam photon-counting laser altimeter significantly improves the sampling density along the orbit while introducing abundant noise. Therefore, noise removal is crucial for the subsequent applications of the photon-counting laser altimeter. This study proposes a new multidimensional statistics algorithm based on the Gaussian mixture model for photon noise removal. The method can simultaneously utilize multiple point cloud statistics and avoid manually selecting denoising thresholds, dividing signals and noise adaptively. To evaluate the performance of the algorithm, it is applied to both simulated ICESat-2 data and real ICESat-2 data, and evaluation indicators including recall, precision, and the harmonic means of recall and precision are used for analysis. Experimental results indicate that compared to denoising using a single statistic, denoising with multidimensional statistics (specifically, the sum of the distance of k-nearest neighbors, the photon density, and the standard deviation of the height difference within an elliptic neighborhood of the target photon) generally yields better results. Multidimensional statistics can influence and constrain each other, leading to improved decision-making outcomes. After comparing the algorithm with mainstream DBSCAN and KNN methods, the results indicate that the GMM method achieves the best denoising results, with average *F*-values 5.22% and 7.26% higher than those of DBSCAN and KNN methods, respectively, across all datasets. In conclusion, this algorithm can comprehensively assess and make decisions based on different statistical measures, demonstrating robustness and adaptability in extracting signal photons, and is worth promoting in the photon denoising applications.**

***Index Terms*—And land elevation satellite-2 (ICESat-2), cloud, gaussian mixture model (GMM), ice, multidimensional statistics, photon denoising, photon-counting lidar.**

Manuscript received 24 March 2024; revised 24 May 2024; accepted 13 July 2024. Date of publication 18 July 2024; date of current version 5 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42071405, Grant 42301445, Grant 42271365, and Grant U22A20566, in part by the National Key R&D Program of China under Grant 2021YFF0704600, in part by the Youth Innovation Promotion Association CAS under Grant 2019130, and in part by the Scientific Research Foundation of Hunan Provincial Education Department under Grant 21B0332. *(Corresponding author: Sheng Nie.)*

Sitong Chen is with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences and the International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: chensitong23@mails.ucas.ac.cn).

Sheng Nie, Xiaohuan Xi, and Cheng Wang are with the International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China, and also with the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: niesheng@aircas.ac.cn; xixh@aircas.ac.cn).

Shaobo Xia is with the Department of Geomatics Engineering, Changsha University of Science and Technology, Changsha 410114, China (e-mail: shaobo.xia@csust.edu.cn).

Feng Zhu is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China, and also with the Hubei Luojia Laboratory, Wuhan 430079, China (e-mail: fzhu@whu.edu.cn).

Xiaoxiao Zhu is with the State Key Laboratory of Earthquake Dynamics, Institute of Geology, China Earthquake Administration, Beijing 100029, China (e-mail: zhuxx@ies.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3430470

I. Introduction

LIGHT detection and ranging (LiDAR) is an active remote sensing detection technology that enables the rapid and precise acquisition of three-dimensional (3-D) spatial information of target objects [1]. The launch of the ice, cloud, and land elevation satellite-2 (ICESat-2) in 2018 [2], equipped with the advanced topographic laser altimeter system (ATLAS), has provided a wealth of photon point cloud data covering the entire globe. These data have been employed in various applications, such as Earth surface elevation extraction, forest height and biomass estimation, and ice sheet and sea ice elevation measurements [3]. ATLAS employs micropulse multibeam photon counting technology, offering small footprints and high sampling density to achieve more accurate 3-D information extraction of Earth's surface [4]. However, since ATLAS records all photon events, it is challenging distinguishing return pulses from target objects from noise generated by atmospheric scattering, solar radiation, and the instrument itself [5], [6]. Thus, an effective noise removal method is required to identify signal photons on target objects accurately.

Existing photon noise removal methods can be divided into four categories: supervised learning-based, unsupervised learning-based, image processing-based, and statistical-based methods.

The supervised learning-based approach involves training classifiers on labeled data samples containing signal and noise, enabling them to classify new data [7]. Classical supervised learning algorithms like decision trees, random forests, support vector machines, and neural networks have been applied to photon noise removal. For instance, random forest models adopt point cloud features, such as height and orbit distance, to construct classifiers that enhance model generalization and provide denoising results [8]. A typical encoder–decoder architecture network, which takes neighboring points of each noise point as input and regresses a displacement vector to push the noise point back to its ground truth position for denoising [9]. Inspired

by neural synapse connections in the brain, the neural network-based method determines neuron weights using a training set. It verifies the classification effect using a testing set of point clouds [10]. A method based on an end-to-end network transforms the input point cloud into the underlying high-dimensional space, and learns the latent manifold of each sampled point to achieve a clean point cloud [11]. However, the main challenge of the supervised methods is that they typically require substantial training datasets. Incomplete or noncomprehensive training datasets can significantly reduce classification accuracy.

Unsupervised learning-based methods primarily involve clustering to classify photon points based on feature similarity or distance [12]. Many algorithms are based on density clustering methods, improving and innovating to suit photon distribution patterns. For example, density-based spatial clustering of applications with noise (DBSCAN) distinguishes signal from noise photons using an elliptical search shape [13]. However, DBSCAN cannot be employed in complex terrains and is sensitive to parameter selection. The method based on the clustering method of ordering points to identify the clustering structure (OPTICS) reduces the sensitivity to parameter, but the number of parameters to be set is large and end up with complicated process [14]. Multilevel asymptotic density clustering methods exhibit good denoising performance across diverse terrains but remain sensitive to input parameters [15]. The local outlier factor algorithm effectively denoises complex terrains with strong background noise based on elliptical density but was computationally intensive [16]. Cluster-based methods lack universality, as a single algorithm cannot meet the requirements of point cloud data processing across different terrains and noise levels, increasing their sensitivity to algorithm parameters.

Image processing-based methods rasterize photon point clouds, distinguish noise and signal pixels, and map the results back to the original point cloud data for classification [17]. Various image processing techniques have been utilized to identify features and extract signal photons, such as the Canny edge detection algorithm, Gaussian filtering, and gradient computation [18]. Median and dimensional filters have also been employed to remove noise photons [19]. The empirical mode decomposition algorithm segments image, and uses the nobuyuki otsu method (OTSU) method to determine the noise-dominated IMFs, and these IMFs were then filtered according to a soft threshold method [20]. The voxel-based spatial filtering method filters noise points using a variable pixel (voxel) binning approach, while retaining the spatial and vertical fidelity of the dataset [21]. Although these image processing-based methods have demonstrated effective noise filtering [22], they convert photon points into raster images, losing vital information, such as the precise position of each photon point. Besides, it is challenging eliminating noise points within the canopy and near the ground.

Statistical-based methods aim to find statistics reflecting photon intrinsic characteristics and set thresholds based on these statistics to differentiate signal from noise photons. For instance, Xia et al. [23] counted the sum of distances from each photon point to the nearest k points using frequency distribution histograms, setting a threshold for denoising. Nie et al. [24] introduced photon density to calculate the number of photon points within an elliptical neighborhood around each photon point as the photon density value for classification. Gwenzi et al. [25] employed elevation as a statistic, assuming a Poisson distribution for the elevation of noise photons. Outliers deviating from this distribution were identified as signal photons. Xie et al. [26] proposed a directional filter method, which achieves the density of the best filter direction by traverse and using the density difference between the point and points in its neighborhood to remove the noise points. Pan et al. [27] innovated a 2-D profile point cloud denoising method based on density and local statistics, which utilizes a statistical outlier removal algorithm to extract the valid signal point cloud. These methods have limitations. They require manual threshold setting, which can be challenging determining, and the optimal threshold varies among datasets, complicating the denoising process. Furthermore, these methods typically employ only a single statistic without utilizing various photon characteristics, leading to unsatisfactory denoising results, especially in complex environments.

Among the four types of denoising methods, the inherent limitations of the supervised learning-based and image processing-based methods are demanding extensive training data and causing information loss, respectively. Algorithm enhancements cannot readily resolve these limitations. Conversely, unsupervised learning-based and statistical-based methods show potential for improvement through innovative algorithm development. Although unsupervised learning-based methods have introduced advanced clustering algorithms, further advancements depend on innovations in machine learning. However, statistical-based methods offer considerable room for improvement by introducing cutting-edge mathematical statistics algorithms.

Consequently, the current work focuses on statistical-based methods and proposes a novel multidimensional statistics denoising algorithm based on Gaussian mixture models (GMM). The algorithm solves two main problems of the existing statistical-based algorithms, including the manual threshold setting and statistics underutilizing. The GMM method converts the classification problem into a probability comparison problem to calculate the probability that each photon point fits the distribution of each single Gaussian model (SGM), and classifies each photon points into the class represented by the SGM with the maximum probability to realize the adaptive selection of the threshold. In order to satisfy the utilization of multiple statistics, one statistic is regarded as 1-D data, and multiple statistics as multidimensional data to put in GMM. The GMM algorithm proposed in this article improves the accuracy of photon denoising algorithms. There are two main advantages of the GMM algorithm that make it highly potential for deeper research in the future. First, the implementation of adaptive classification eliminates the need to determine the threshold to noise point removal through human experience, making the basis of classification more scientifically grounded and reducing the pressure of manual settings. Second, the use of multidimensional statistics provides the GMM algorithm with significant room for innovation. There are many statistics related to point clouds, each with different emphases. The GMM algorithm makes it possible to denoise using multiple different statistics simultaneously. More effective and complementary statistics discovered in

future research can all be used as inputs to the GMM algorithm to achieve more precise denoising effects. The following work are performed to prove the advantages of our proposed algorithm.

1) Validation of the denoising effectiveness of our novel GMM algorithm using both simulated and real ICESat-2 datasets, showcasing its robustness and generalizability in handling photon cloud data with varying noise ratios, acquisition periods, and terrains.
2) Demonstration of the superiority of the GMM algorithm's denoising approach based on multidimensional statistical quantities over denoising methods using individual statistical quantities, addressing the limitations of poor applicability of single statistical quantities and significant variations in denoising results across different datasets.
3) Improvement in denoising efficacy compared to unsupervised DBSCAN and KNN algorithms, along with the algorithm's adaptive implementation of point cloud classification to avoid the instability and empirical reliance associated with manually setting classification thresholds in previous approaches. The strong denoising performance of our algorithm provides a solid foundation for further extraction and interpretation of signal point cloud information in practical applications.

The rest of this article is arranged as follows. The datasets are introduced in Section II. The principles and processes of our proposed algorithm are elucidated in Section III. Section IV analyzes and evaluates the denoising results. Finally, Section V concludes this article.

II. Dataset

A. Simulated ICESat-2 Data

NASA acquired simulated ICESat-2 data to assess photon data quality and performance before the ICESat-2 launch. This data is freely accessible at Airborne Data | ICESat-2 (nasa.gov), NASA conducted random sampling to achieve three different noise rates: 0.5, 2, and 5 MHz [4]. Our study utilized data from Pine Barren regions in New Jersey, USA (Cedar2 and Cedar4 datasets) and the Smithsonian Environmental Research Center in Maryland, USA (SERC1, SERC3, and SERC5 datasets).

Simulated ICESat-2 data was labeled with signal and noise photons, enabling precise denoising and quantitative evaluation. Fig. 1 describes the photons obtained from Cedar2 data with different noise rates. Fig. 2 displays photons from five datasets with a 0.5 MHz noise rate.

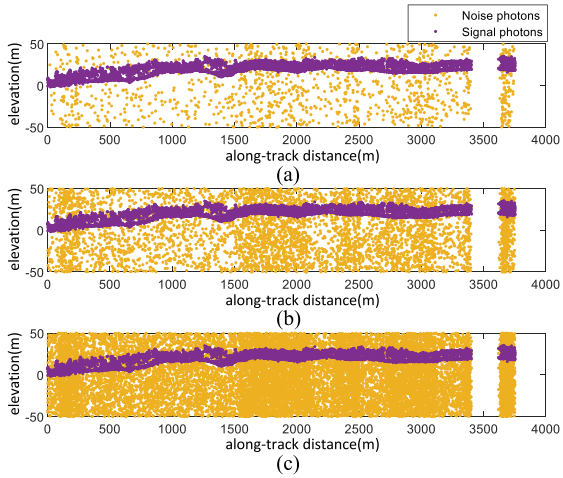


Fig. 1. Signal and noise photons of Cedar2 dataset with different noise rates. (a) Cedar2 (0.5 MHz). (b) Cedar2 (2 MHz). (c) Cedar2 (5 MHz).

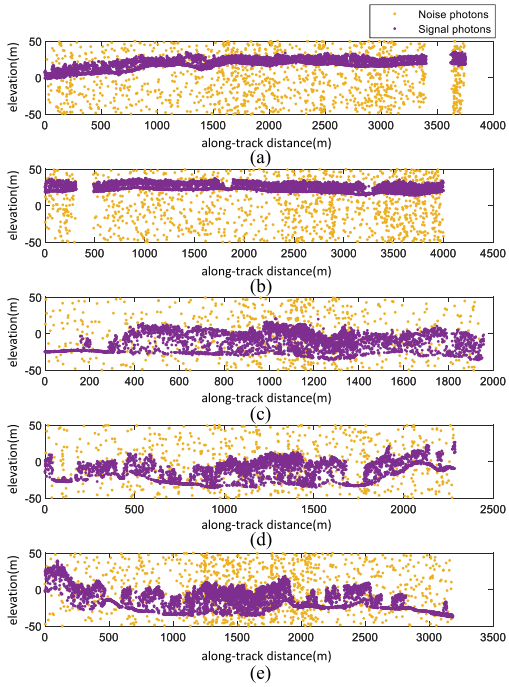


Fig. 2. Signal and noise photons of five datasets with a 0.5 MHz noise rate. (a) Cedar2 (0.5 MHz). (b) Cedar4 (0.5 MHz). (c) SERC1 (0.5 MHz). (d) SERC3 (0.5 MHz). (e) SERC5 (0.5 MHz).

B. Real ICESat-2 Data

Real ICESat-2 data assessed the proposed algorithm's performance comprehensively, providing more topographic features and data acquisition conditions. As a new-generation satellite-borne LiDAR, ICESat-2 collects data using the ATLAS system, contributing to global environmental research [28]. It aids in quantifying the impact of polar ice sheets on sea-level change, optimizing ice sheet prediction models, estimating ice thickness, and measuring vegetation height.

Real ICESat-2 data was acquired using the ATLAS system, comprising lasers, a diffractive splitter, a laser alignment system, and a star sensor. Table I presents detailed information on ICESat-2 and ATLAS. The diffractive splitter divides the laser beam into six beams, organized into three parallel groups along the rail direction, as depicted in Fig. 3. In each group, the distance between two laser beams is approximately 90 m, and adjacent groups are separated by about 3.3 km.

NASA provides 21 ATLAS data products labeled ATL00 to ATL21 (excluding ATL05), categorized into Level 0, Level 1, Level 2, and Level 3. These products can be freely downloaded from the National Snow and Ice Data Center's official website[4].[1]

[1]Online. [Available]: ICESat-2(nsidc.org)

TABLE I
ICESAT-2/ATLAS PARAMETERS

ICESat-2		ATLAS	
Parameter	Value	Parameter	Value
Altitude in Orbit	500 km	Transmitting Frequency	10 kHz
		Laser wavelength	532 nm
		Footprint diameter	~ 12 m
Orbital Inclination	92°	Footprint spacing along track	about 0.7 m
		Pulse width	1.5 ns
Coverage Area	88°S–88°N	Pulse intensity	0.2–1.2 mJ
		Beam intensity (strong)	175±17 μJ
Repetition Cycle	91 days	Beam intensity (weak)	45±5 μJ

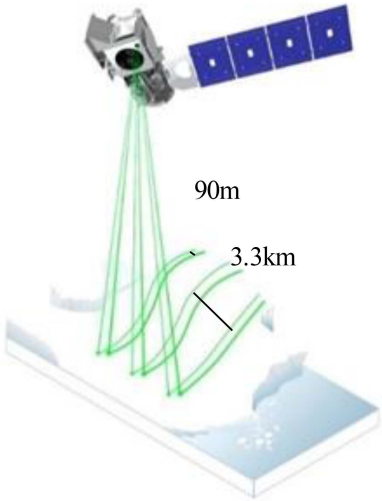


Fig. 3. Distribution of the ATLAS laser beams [1].

This study employed the ATL03 product, containing information on the longitude, latitude, and elevation of photon points. Our study incorporated four datasets, including two daytime and two nighttime datasets, acquired in 2019 and 2020, all with manually labeled true values, each comprising three data strips. The real datasets include diverse land cover types and extensive areas and does not specify a noise rate. The smallest group contains 7000 points, and the largest has 130 000 points. Figs. 4 and 5 present one data from each group for display purposes.

III. METHODOLOGY

The denoising process consists of four main components: photon point cloud preprocessing, photon point features extraction, denoising method based on the Gaussian Mixture Model, and residual noise removal. To provide a comprehensive overview of the entire process, a flowchart is presented in Fig. 6.

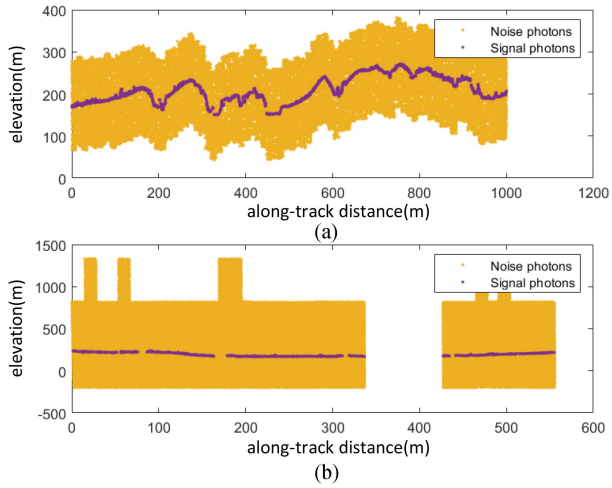


Fig. 4. Signal and noise photons of daytime datasets. (a) Day 1. (b) Day 2.

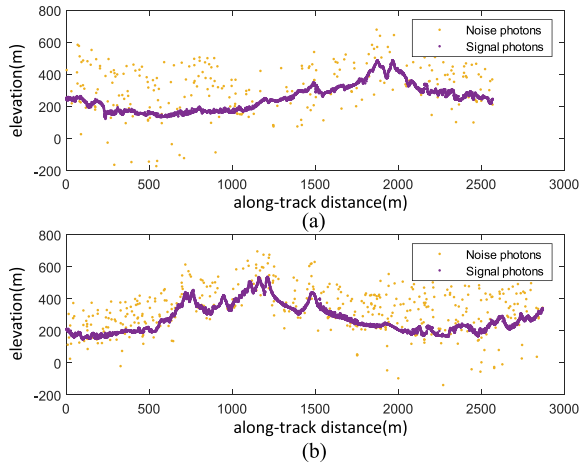


Fig. 5. Signal and noise photons of nighttime datasets. (a) Night 1. (b) Night 2.

A. Photon Point Cloud Preprocessing

As mentioned in Section II, simulated ICESat-2 data displayed dispersed point clouds across the 2-D elevation profile. Conversely, real ICESat-2 data showed uneven distribution with prominent areas, as indicated in Fig. 7. To mitigate the misclassification of photons in the step of GMM due to substantial variations in statistics between prominent and other areas, the grid method is introduced to remove noticeable noise areas.

The grid method divides the 2-D point cloud into grids along the elevation and track directions. It identifies the grid containing the most photons within the same track region (called target grid), takes the region above and below the corresponding elevation direction of the grid as a buffer, and designates photons in this region as signal photons. Due to the relatively small elevation variations in the real ICESat-2 dataset, with most not exceeding 50 m, the buffer size is set at twice this value as a precaution, which means 100 m in total elevation range for the target grid—50 m above and below. Additionally, the grid size along the trajectory direction can be set to a larger value (100 m) as it has less impact on the identification of the target

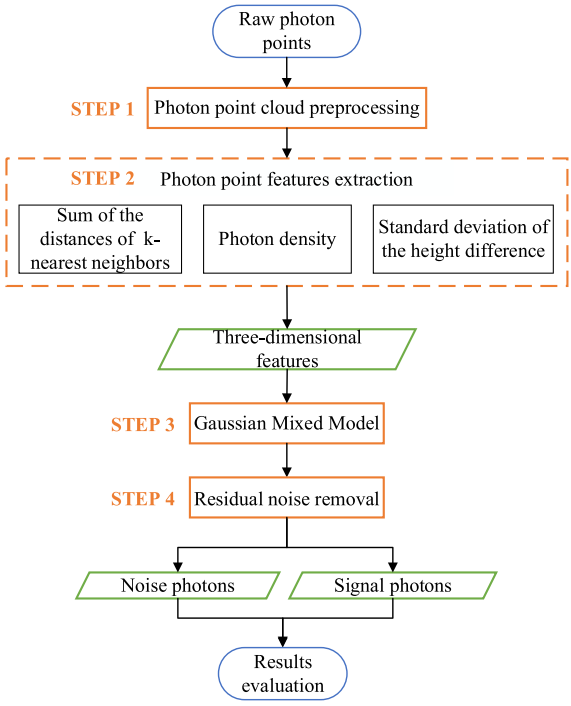


Fig. 6. Flowchart of the denoising process.

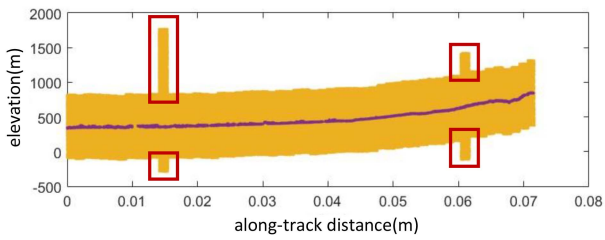


Fig. 7. Prominent areas of a daytime dataset.

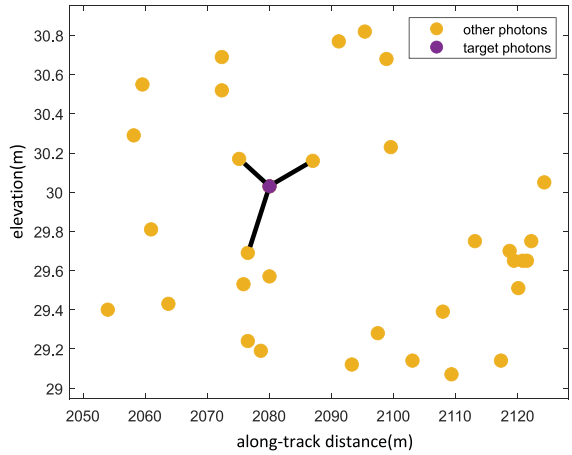


Fig. 8. Schematic diagram of the distance of k-nearest neighbors.

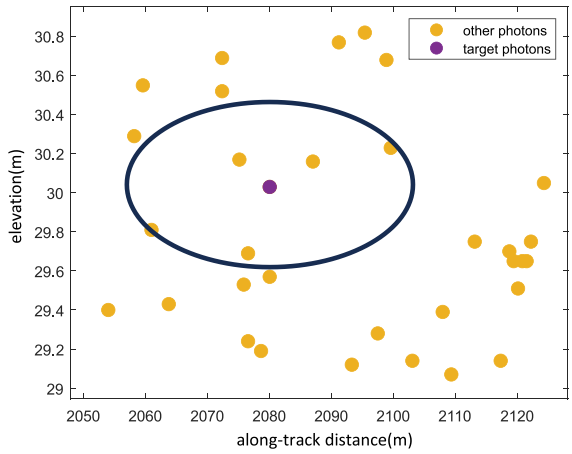


Fig. 9. Schematic diagram of the photon density.

grid. The width of the grid in the elevation direction needs to be much smaller to distinctly differentiate areas where signals are concentrated and accurately locate the target grid. Therefore, the grid size in the elevation direction is set at 1/10 of 50 m, that is 5 m [15], [30]. The selection of grid size and the value of the above and below region is quite flexible, and do not need to be constrained to specific optimal values. As long as the grid size can roughly identify where more signal occurs, and the buffer zone can contain all signals and excludes prominent areas, the values are effective.

B. Photon Point Features Extraction

The more significant the difference in the features for distinguishing signal from noise, the better the distinction. The distribution in the 2-D profiles is considered the core difference between signal and noise. The number of signal photons is much larger than noise photons. More importantly, the noise generation process reflects a random noise distribution while the signal distribution is clustered. In this case, satisfactory results can be obtained using appropriate features that can reflect the difference in signal distribution and noise distribution.

Three-dimensional features are selected for the GMM, including the sum of the distances of k-nearest neighbors, the photon density, and the standard deviation of height differences within an elliptical neighborhood (HDSTD). They determine the difference in signal and noise photons from three aspects: distance, density, and elevation concentration, which are the standard most apparent features to distinguish the difference in distribution.

1) Distance-Based Feature: The sum of distances to a certain number of the nearest photons for signal photons is smaller than noise photons, reflecting a good classification effect. Fig. 8 illustrates the calculation process of the sum of the distance of k-nearest neighbors.

2) Density-Based Feature: Photon density quantifies the number of photons within the elliptical neighborhood around the target photon, indicating that higher photon density is more likely to indicate a signal photon. An elliptical neighborhood shape is employed since the photon density along the orbital direction is generally higher than along the elevation direction [31] (as shown in Fig. 9). The elliptical neighborhood's target and edge photons contribute to photon density.

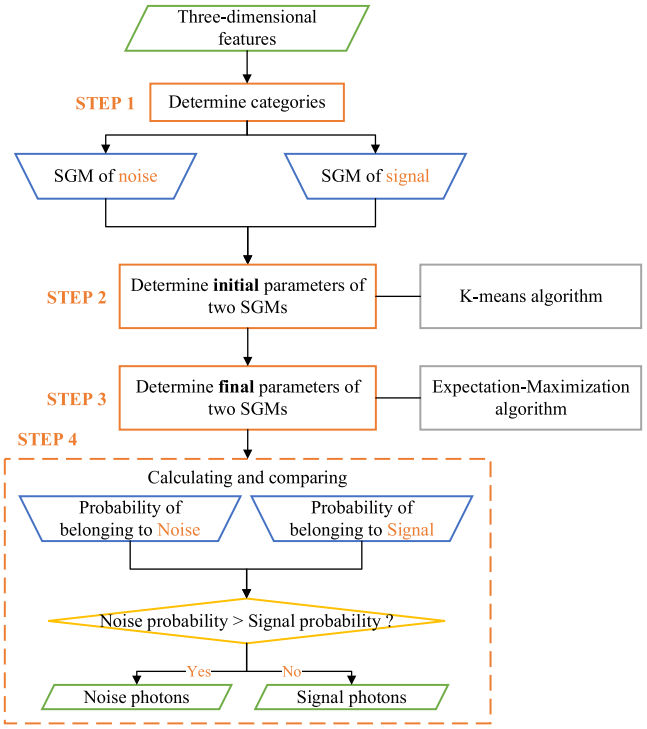


Fig. 10. Flowchart of the GMM algorithm.

3) Height-Based Features: These features include height difference, height standard deviation, and the HDSTD. Among them, HDSTD can better reflect the height closeness between the target point and other points in its neighborhood. Since the signal points are mainly gathered on the elevation, while the noise points are mainly random, the two have different HDSTD values, making denoising effective

$$\mathrm{HDSTD} = \sqrt{\frac{\sum_{i=1}^{n} (z_i - z_{\mathrm{ave}})^2}{n-1}} \tag{1}$$

where z_i is the height of the ith photon, z_{ave} is the average height of the photons in the elliptic neighborhood, and n is the photon density value.

The above three statistics reflect the difference in signal and noise points with different denoising emphases. However, a single statistic cannot be perfectly applied to all situations. The GMM model achieves a better denoising effect when dealing with various data types by simultaneously utilizing several statistics. The selection of the parameter values in the above three statistics takes into consideration both noise reduction effectiveness and algorithm timeliness. Among them, the major and minor axes of the ellipse are set as 6 and 2 m, and the number of the nearest neighbors is set to 10. As evidenced in Section IV-C, the selection of these values meet the aforementioned two requirements.

C. Denoising Method Based on the Gaussian Mixture Model

The GMM algorithm dynamically partitions the input features into predefined categories, enabling the separation of signal photons and noise photons in point cloud denoising. The GMM algorithm involves stages of parameter initialization using the k-means algorithm, final parameter values determination of single Gaussian models utilizes expectation-maximization (EM) algorithm, and procedures for probability classification. The flowchart illustrating these steps is presented in Fig. 10.

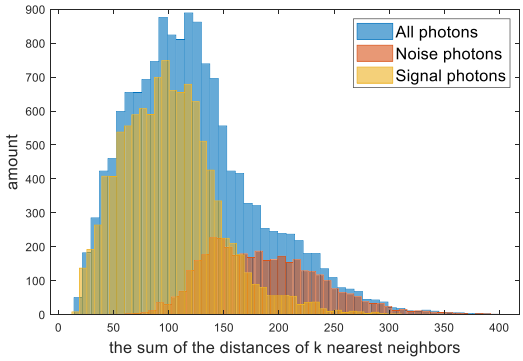


Fig. 11. Statistical histograms of the mixed Gaussian distribution (blue) and single Gaussian distribution (yellow and red) of the sum of the distance of k-nearest neighbors of the simulated ICESat-2 data.

GMM treats input data as distinct groups following single Gaussian distributions with a predefined number of groups [14]. For the photon denoising domain, the value of the groups is determined by the number of categories to be classified. Since we need to divide all photons into two categories: signal photons and noise photons, the value of Gaussian components is set to 2, which means the statistical histogram corresponding to each feature approximately follows a mixed Gaussian distribution comprising two single Gaussian distributions (Fig. 11 illustrates the distribution of the sum of distances of k-nearest neighbors using simulated ICESat-2 data as an example). However, a single statistic of signal photons and noise photons may not have distinctive discriminative features, leading to the limited denoising effect. Therefore, multiple statistics are considered multidimensional information input to the GMM, to provide more comprehensive information and achieve better decision-making for photon cloud denoising. The input data can be divided into signal and noise photons by separating the two single Gaussian distributions in the mixed Gaussian distribution.

The core of the GMM algorithm is to estimate the parameters of single Gaussian distributions for both groups, including weight, mean and variance matrices. The weight determines the contribution of the SGM distribution to the overall GMM distribution, representing the extent to which a statistic influences the overall decision-making process. The mean and variance matrices determine the shape of the SGM, including the central position and the width of the distribution, as illustrated in Fig. 12. The estimation of these parameters can be divided into two steps: initial value determination and final value acquisition. In initial value determination step, the k-means algorithm is applied to obtain the initial value of the three parameters. In final value acquisition step, the EM algorithm is used for iterating to get the final exact parameter value. The EM algorithm involves E-step and M-step. Repeat these two steps until the stopping condition is met, which means that the change in parameters is below a threshold or the predefined number of iterations is reached.

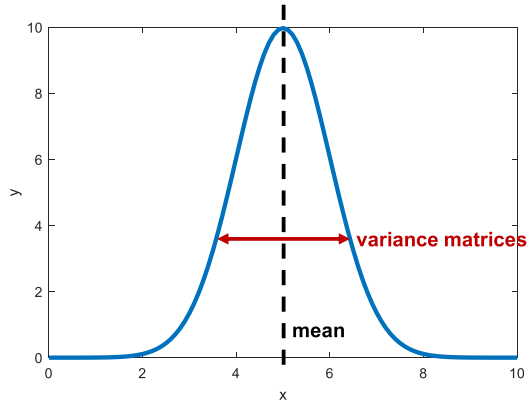


Fig. 12. Shape of the SGM determined by the mean and variance matrices.

After all the parameters of the two SGM distributions are determined, the probability of each photon belonging to these two distributions can be computed. A photon with a higher probability of belonging to the signal SGM is distinguished as a signal photon. In contrast, a photon with a higher probability of belonging to the noise SGM is distinguished as noise. In order to illustrate the classification of using multi-dimensional statistics visually, data originated from three distinct groups distributed in a 3-D space as an example. Before classification, the data may appear scattered and disordered. However, after classification, it converges to three ellipsoidal shapes, which is compatible with the characteristics of Gaussian distributions (as depicted in Fig. 13).

Mathematically, the probability distribution density function ($p(x_i)$) for this dataset can be represented with the following weighted function:

$$p(x_i) = \sum_{j=1}^{M} \alpha_j N_j \left(x_i, \mu_j, \Sigma_j\right) \tag{2}$$

where i is the serial number of the sample point x_i, M is the number of SGM, j is the serial number of the SGM, α_j denotes the weight of the jth SGM, and $N_j(x_i, \mu_j, \Sigma_j)$ is the probability density function (PDF) of the jth SGM, described as follows:

$$N_j \left(x_i, \mu_j, \Sigma_j\right) = \frac{1}{\sqrt{(2\pi)^{|\Sigma_j|}}} \exp\left[-\frac{1}{2}\left(x_j - \mu_j\right)\Sigma_j^{-1}\left(x_j - \mu_j\right)\right] \tag{3}$$

where x_j denotes the sample vector (column vector), μ_j represents the model expectation, and Σ_j corresponds to the model variance.

Two SGMs are considered for a photon point cloud dataset: one for signal and one for noise. Each SGM comprises three unknown parameters. The parameters for the ith SGM are defined as follows:

$$\theta_i = \left(\alpha_i, \mu_i, \Sigma_i\right). \tag{4}$$

The unknown parameters required for the GMM in the photon denoising process are as follows:

$$\Theta = \left(\theta_1, \theta_2\right)^{\mathrm{T}} \tag{5}$$

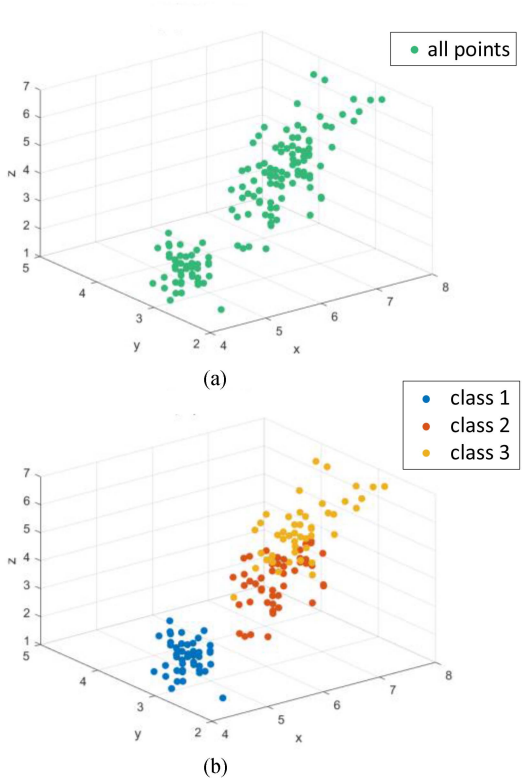


Fig. 13. Data corresponding to a Gaussian mixed distribution. (a) Unclassified. (b) Classified.

The input features of the photon point cloud data can be employed to estimate these unknown parameters. The estimation process involves two stages: initializing initial values and refining them to obtain the final values iteratively.

1) Determining the Initial Parameters: There are two methods to determine the initial values of the unknown parameters. The first method directly assigns values to the unknown parameters, sets the covariance matrix Σ_{i0} as the unit matrix, defines the proportion of the two data groups as $\alpha_{i0} = 1/2$, and sets the mean value μ_{i0} as a random number. The second method clusters the data into two groups using the k-means algorithm to calculate the mean value μ_{i0}, the proportion α_{i0}, and the covariance matrix Σ_{i0} of the two groups. Since the latter method provides estimated values of the unknown parameters closer to their final values, the number of iterations is reduced, and efficiency is improved. The k-means clustering method estimates the initial values of unknown parameters.

The core concept of the k-means clustering method is as follows: it involves setting the value of k and initializing the k-cluster center points. Subsequently, the distance between each photon points and the center point of each cluster is calculated, assigning each photon point to the class of the nearest cluster center point. Since the photon points are assigned, the center of each cluster is recalculated [32]. The process of photon points assignment and cluster center point update is repeated until the change in cluster center point coordinates becomes smaller than a predefined threshold value or reaches the maximum

allowable number of iterations. At this point, the parameters are determined conclusively [33].

The mathematical expression for this process can be represented as follows: The feature dataset X of photon point cloud contains n photon points, where $X_i$in the following equation represents the 3-D properties corresponding to the *i*th photon point

$$X = \{X_1, X_2, \ldots, X_n\}. \tag{6}$$

Distance calculation requires initializing the *k* cluster center points obtained through random selection. Two clusters are employed for the photon denoising process, expressed as follows:

$$C = \{C_1, C_2\} \tag{7}$$

where C_j represents the 3-D properties of the *j*th cluster center point.

The Euclidean distance between each photon point and the cluster center points can be calculated as follows:

$$\mathrm{dis}(X_i, C_j) = \sqrt{\sum_{h=1}^{3} (X_{ih} - C_{jh})^2} \tag{8}$$

where X_{ih} represents the *h*th property of the *i*th photon point, and C_{jh} represents the *h*th property of the *j*th cluster center point.

After classifying each photon point into the cluster closest to its center point, the following two clusters are obtained:

$$S = \{S_1, S_2\}. \tag{9}$$

By calculating the mean value of each dimension within these two clusters, the following two new cluster center points are obtained iteratively:

$$C_j = \frac{\sum X_{i \in S_j} X_i}{n}. \tag{10}$$

The iteration continues until the new cluster center point closely matches the previous one, while the difference between the two becomes lower than the threshold value. At this point, the iteration process stops, and the initial values of the unknown parameters Θ are determined.

2) Obtaining the Final Parameters: Once the initial values of the unknown parameters have been determined, the final parameter values (Θ) are estimated using the EM algorithm. This adaptive parameter estimation is a key feature of our approach, highlighting the convenience and rationality of the GMM. The EM algorithm comprises the expectation step (E-step) and the maximization step (M-step).

a) E-step: Let β_j be the posterior probability of α_j, expressed as follows:

$$\begin{aligned} \beta_j &= E(\alpha_j | x_i; \Theta) \\ &= \frac{\alpha_j N_j(x_i; \Theta)}{\sum_l^M \alpha_l N_l(x_i; \Theta)}, 1 \le i \le n, 1 \le j \le M \end{aligned} \tag{11}$$

where N_j represents the PDF of the *j*th SGM calculated using formula (3), the first Θ is obtained via the *k*-means algorithm, while the M-step calculates the other Θ values.

b) M-step: In this step, all values to be estimated in Θ are updated using β_j obtained in the E-step, following these formulas:

$$\alpha_j{}' = \frac{\sum_{i=1}^{N} \beta_{ij}}{N} \tag{12}$$

$$\mu_j{}' = \frac{\sum_{i=1}^{N} \beta_{ij} x_i}{\sum_{i=1}^{N} \beta_{ij}} \tag{13}$$

$$\Sigma_j{}' = \frac{\sum_{i=1}^{N} \beta_{ij} (x_i - \mu_j{}')(x_i - \mu_j{}')^{\mathrm{T}}}{\sum_{i=1}^{N} \beta_{ij}}. \tag{14}$$

c) Convergence condition: Let $\theta_i{}' = (\alpha_i{}', \mu_i{}', \Sigma_i{}')$ be the parameters after each M-step, and $\theta_i = (\alpha_i, \mu_i, \Sigma_i)$ be the parameters before the update. The iteration process is halted when the parameter change falls below the predefined threshold or the maximum number of iterations is reached. Specifically, the decision-making threshold is set at 10^{-10}, and the criterion for stopping the loop is defined as follows:

$$\begin{cases} \frac{\Theta}{\Theta'} - 1 < \varepsilon \\ n \ge n_m \end{cases} \tag{15}$$

where ε represents the threshold value set to 10^{-10}, and n_m is the upper limit set to 1000.

The probability of each photon point belonging to either of the two groups can be calculated upon determining the parameters of the two single Gaussian models through the EM algorithm. By comparing these probability values, each photon point can be classified into a cluster with a higher probability.

D. Residual Noise Removal

A substantial portion of the noise photons is successfully removed using the GMM algorithm. However, a fraction of remaining noise photons comprises points situated considerably from the signal photons. In such cases, a three-sigma rule-based approach is employed to enhance the denoising process.

This approach utilizes the sum of distances between each remaining point and its k-nearest neighbors as a group of statistical data to obtain the mean value denoted as μ and the standard deviation represented by σ. Since the greater the sum of distances, the lower the probability that a given photon point corresponds to a signal point. It is reasonable to assume that the statistical distribution of photon point is approximately normal. For data that follow a normal distribution, the percentage of data points within three standard deviations from the mean are fixed at 99.73%. This indicates that the probability of the sum of distances between each remaining point and its k-nearest neighbors exceeding ε is very low, allowing it to be classified as an outlier in the data. Since the distance values of most signal points are close, these outliers are more likely the residual noise points that were not excluded by the algorithm. Therefore, we set $\mu + 3\sigma$ as the threshold for the sum of distances. Photon points whose sum of distances exceeds this threshold are classified as noise and consequently removed. The following decision law is established according to the three-sigma rule:

$$\begin{cases} \varepsilon < \mu + 3\sigma & \text{signal photon} \\ \varepsilon >= \mu + 3\sigma & \text{noise photon} \end{cases} \tag{16}$$

where ε is the threshold value.

E. Evaluation and Method Comparison

Our study employs qualitative and quantitative evaluation methods to analyze and assess noise reduction outcomes comprehensively. For qualitative evaluation, two sets of data are considered. First, the denoising process is illustrated using the actual daytime data, showcasing the effectiveness of the three-step approach, including the preprocessing with the grid method, Gaussian Mixture Model denoising, and residual noise removal. Second, the effectiveness of multidimensional statistics is compared with the three individual statistics using the simulated data.

In quantitative evaluation, three statistical evaluation metrics are chosen: recall (R), precision (P), and F-value, represented by the following equations:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{17}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{18}$$

$$F = \frac{2P \times R}{P + R} \tag{19}$$

where TP represents the number of correctly classified real signal photons, FN indicates the number of actual signal photons misclassified as noise, and FP signifies the number of noise photons incorrectly categorized as signal photons.

The denoising results of the multidimensional GMM method are compared with those single statistics approach before the residual noise removal to assess their effectiveness. Three types of statistics are separately applied within the Gaussian mixture model to obtain the individual statistics results. Additionally, the denoising results of the proposed algorithm are compared with two mainstream types of denoising algorithms, including the DBSCAN [13] and KNN methods [29], to demonstrate the superiority of our algorithm.

IV. Results and Discussion

A. Qualitative Evaluation and Analysis

In order to illustrate the effect of preprocessing, the GMM method, and residual noise removal, a representative dataset was selected from a daytime dataset. The results are presented in the following figures. Fig. 14 demonstrates the excellent performance of the grid method in the preprocessing process. In a dataset with such dense point clouds, it is challenging to visually discern the concentrated signal areas. However, through the preprocessing step that compares the number of points in different height intervals along the trajectory span uniformly, the preprocessing method accurately identifies the height interval where the signal point clouds are located, effectively removing the photon clouds far from the signal. Fig. 15 illustrates the processing effect of the GMM algorithm. It can be observed that after being processed by the GMM algorithm, the majority of residual noise has been identified and removed, revealing the overall outline of the signal photons, with only a small portion of discrete noise points remaining to be eliminated.

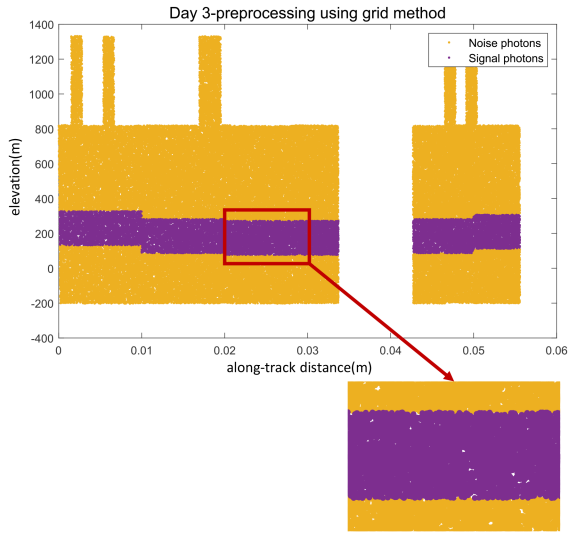


Fig. 14. Denoising effect of the grid method-based preprocessing.

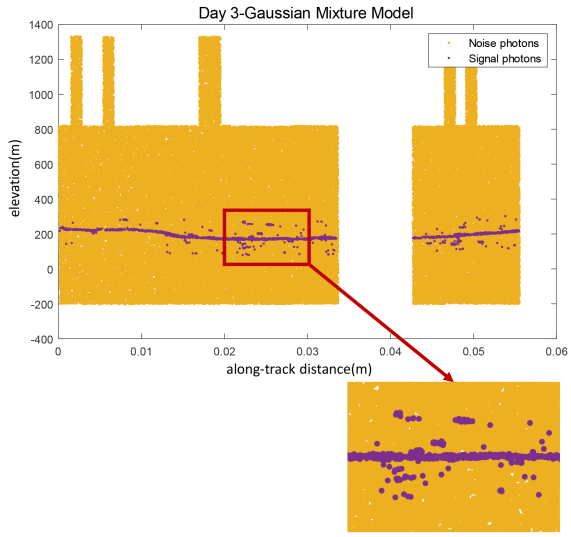


Fig. 15. Denoising effect of the Gaussian Mixture Model.

Fig. 16 shows the results of the residual noise removal step utilizing the three-sigma rule after eliminating residual noise. A noticeable comparison with Fig. 14 reveals that most of the remaining discrete noise points have been identified and removed by the method. Among these steps, the core contribution of the GMM method in this study is the most significant and impactful. Through the sequential contributions of these three steps, noise in the point cloud is eliminated step by step successfully, ultimately achieving the desired results.

In order to demonstrate the superiority of multidimensional statistics to single statistics, a representative dataset was selected from a simulated dataset for display. The reference results are shown in Fig. 17, while the denoising effect of inputting the three statistics as multidimensional statistics into the Gaussian mixture model is shown in Fig. 18, while the denoising effect of inputting the three statistics separately into the Gaussian mixture model is shown in Fig. 19.

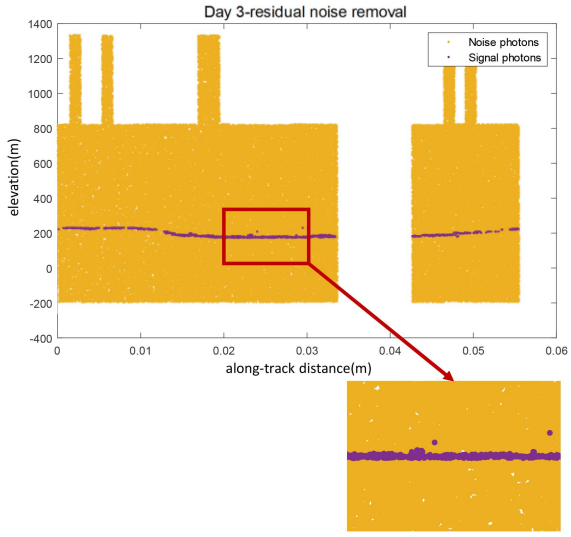


Fig. 16. Denoising effect of residual noise removal.

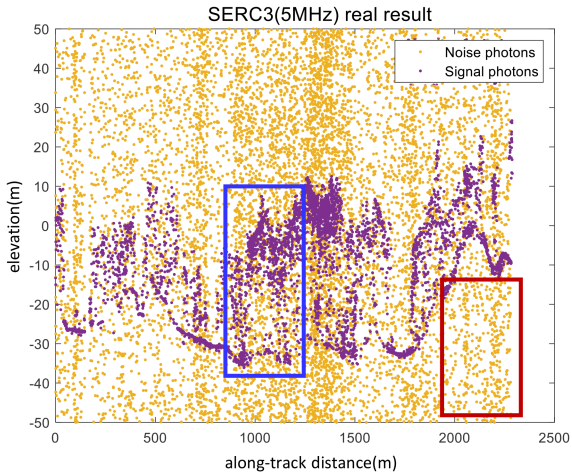


Fig. 17. Real result of a simulated ICESat-2 data.

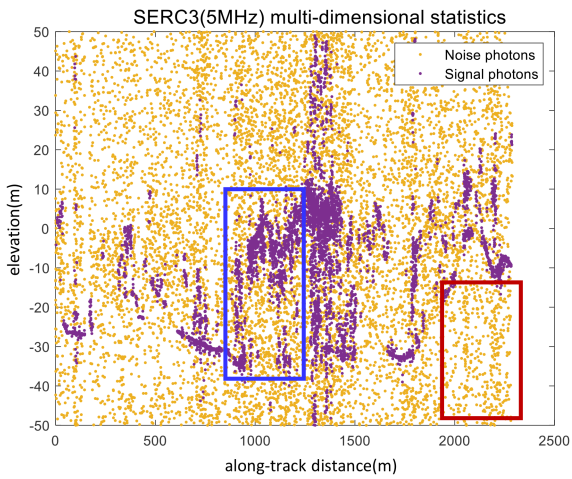


Fig. 18. Denoising effect of multidimensional statistics.

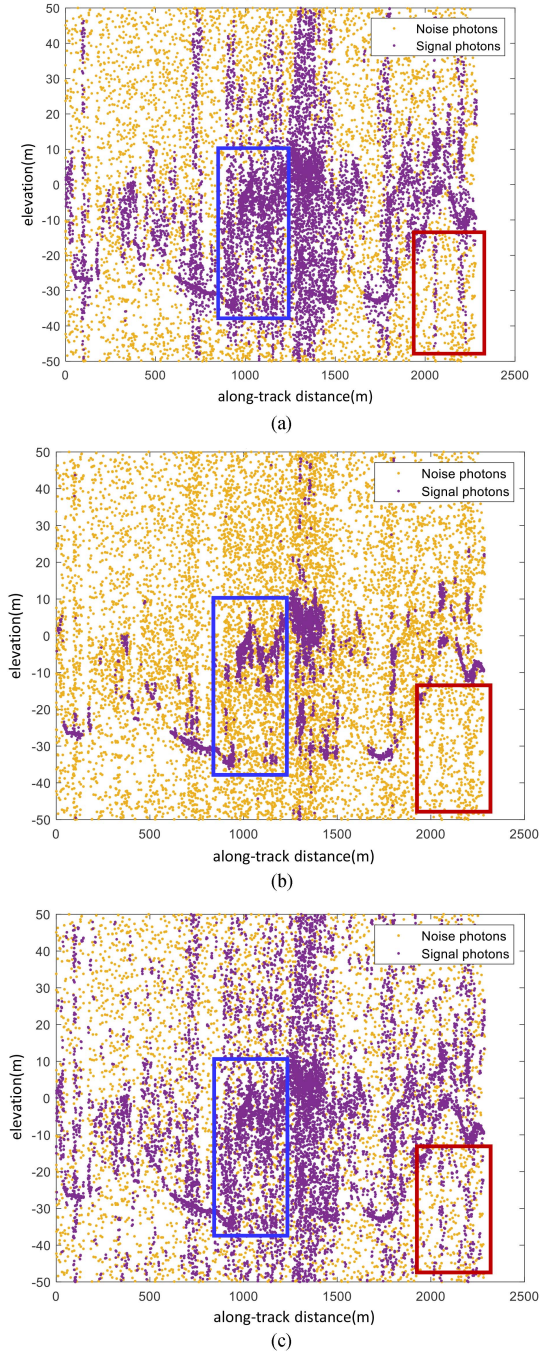


Fig. 19. Denoising effect of the three single statistics. (a) SERC3 (5 MHZ) the sum of distance of k nearest neighbors. (b) SERC3 (5 MHZ) photon density. (c) SERC3 (5 MHZ) HDSTD.

Upon comparing the three figures, it is apparent that the denoising effect of the multidimensional statistics in Fig. 18 is the closest to the reference result in Fig. 17. However, each of the individual denoising results of the three statistics exhibits its own respective drawbacks. Many noise photons are incorrectly identified as signals in the denoising results for both the sum of the distance of k-nearest neighbors and the HDSTD methods [as shown in the red rectangle box in Fig. 19(a) and (c)]. At the same time, the photon density results miss many signals [as shown in the blue rectangle box in Fig. 19(b)].

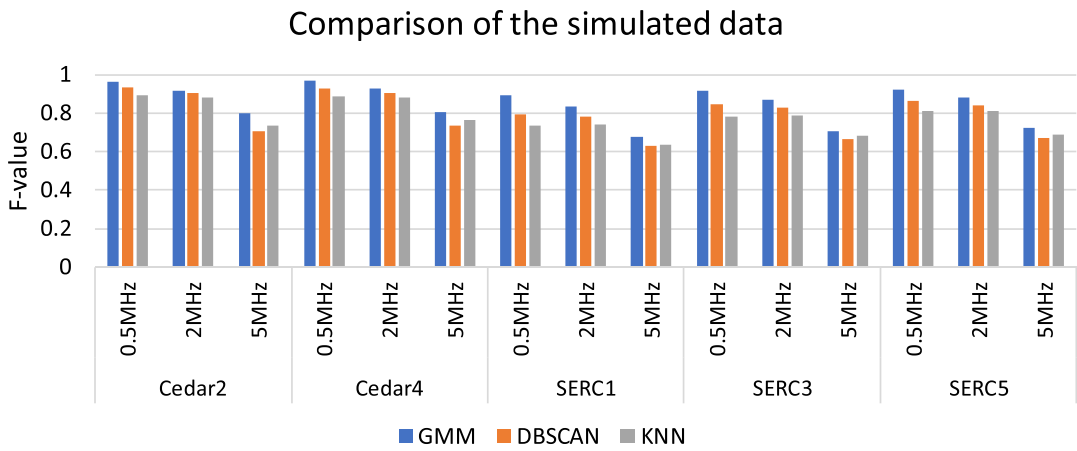


Fig. 20. Comparison of the *F*-values of the GMM, DBSCAN, and KNN methods in simulated data.

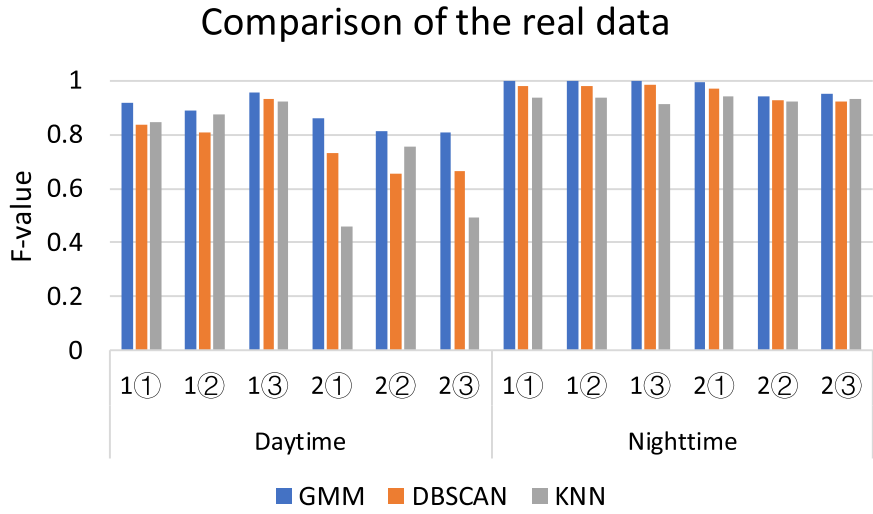


Fig. 21. Comparison of the *F*-values of the GMM, DBSCAN, and KNN methods in real data.

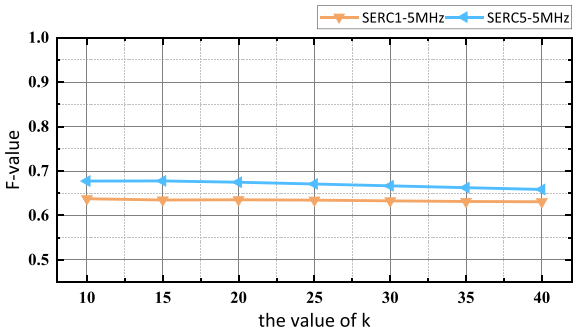


Fig. 22. Influence of *k* on the denoising effect of *k*-nearest neighbors.

In this context, the GMM-based multidimensional statistics denoising algorithm not only preserves more signals but also removes more noise, successfully combining the strengths of these methods, enhancing both the correct signal and noise recognition rates, and achieving superior denoising results.

B. Quantitative Evaluation and Analysis

The simulated and real ICESat-2 data demonstrate the robustness and self-adaptability of the GMM method. It typically yields the best denoising results when all statistics are effective and achieves reasonably good results even when some statistics are negative. Along with the advantages of adapting classification without the need for manual threshold selection, the GMM method is worth promoting.

Table II shows the denoising results of multidimensional statistics and the three single statistics using the GMM method of simulated ICESat-2 data. It reveals that the GMM method consistently achieves the best denoising results, especially in datasets with high noise rates. Although the GMM method may not outperform the other methods in low noise rate datasets, it still delivers commendable results. Despite the challenges posed by photon density values, the GMM algorithm provides denoising results that closely match the best among the four methods, indicating the robustness of the GMM approach.

The denoising results for the *k*-nearest neighbors, photon density, and HDSTD methods are worse in high noise rate datasets due to different reasons. Photon density exhibits a high *P*-value but low *R*-value, indicating effective noise removal, while a significant proportion of true signal photons are missed. Conversely, the HDSTD method demonstrates a high *R*-value but low *P*-value, signifying the successful identification of most

TABLE II
DENOISING RESULTS AFTER APPLYING THE GMM METHOD TO SIMULATED ICESAT-2 DATA

Datasets	Noise Rate	GMM			K Nearest Neighbors			Photon Density			HDSTD		
		R	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
Cedar2	0.5 MHz	0.949	0.977	0.963	0.970	0.969	0.970	0.361	0.990	**0.529**	0.948	0.977	0.962
	2 MHz	0.912	0.917	**0.915**	0.928	0.896	0.911	0.915	0.906	0.911	0.956	0.863	0.907
	5 MHz	0.844	0.726	**0.781**	0.916	0.626	0.744	0.806	0.752	0.778	0.976	0.453	**0.619**
Cedar4	0.5 MHz	0.953	0.979	0.966	0.988	0.971	0.979	0.342	0.991	**0.509**	0.956	0.978	0.967
	2 MHz	0.943	0.916	**0.929**	0.958	0.899	0.927	0.927	0.920	0.924	0.964	0.878	0.919
	5 MHz	0.895	0.711	**0.793**	0.937	0.642	0.762	0.797	0.777	0.787	0.983	0.491	**0.655**
SERC1	0.5 MHz	0.826	0.959	0.888	0.877	0.952	0.913	0.461	0.980	**0.627**	0.825	0.959	0.887
	2 MHz	0.866	0.809	0.836	0.898	0.807	0.850	0.572	0.891	**0.696**	0.865	0.809	0.836
	5 MHz	0.698	0.587	**0.638**	0.857	0.504	0.635	0.601	0.631	0.616	0.927	0.437	**0.594**
SERC3	0.5 MHz	0.869	0.975	0.919	0.917	0.964	0.940	0.470	0.989	**0.638**	0.868	0.975	0.919
	2 MHz	0.899	0.842	**0.870**	0.894	0.844	0.868	0.597	0.934	0.729	0.898	0.840	0.868
	5 MHz	0.744	0.668	**0.704**	0.938	0.481	0.636	0.620	0.754	0.681	0.947	0.449	**0.609**
SERC5	0.5 MHz	0.884	0.962	0.921	0.939	0.952	0.945	0.464	0.982	**0.631**	0.883	0.962	0.921
	2 MHz	0.908	0.842	**0.874**	0.884	0.860	0.872	0.628	0.929	0.750	0.907	0.840	0.872
	5 MHz	0.759	0.672	**0.713**	0.896	0.540	0.674	0.655	0.757	0.702	0.950	0.452	**0.613**

Red bold values means indicates poor results, and the green bold values indicates ideally results.

TABLE III
DENOISING RESULTS AFTER APPLYING THE GMM METHOD TO REAL ICESAT-2 DATA

Datasets	Number	GMM			K Nearest Neighbors			Photon Density			HDSTD		
		R	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
Daytime	1	0.889	0.942	**0.915**	0.937	0.936	0.937	0.846	0.907	0.876	0.939	0.508	**0.660**
		0.805	0.991	0.888	0.961	0.960	0.960	0.797	0.985	0.881	0.927	0.683	**0.786**
		0.923	0.985	**0.953**	0.980	0.970	0.975	0.923	0.976	0.949	0.975	0.708	**0.820**
	2	0.843	0.748	**0.793**	0.822	0.374	0.514	0.825	0.812	0.817	0.427	0.447	**0.437**
		0.872	0.717	**0.788**	0.841	0.341	0.486	0.851	0.760	0.803	0.510	0.403	**0.450**
		0.704	0.792	**0.746**	0.492	0.961	0.651	0.686	0.838	0.754	0.405	0.554	**0.468**
Nighttime	1	0.998	0.999	**0.999**	0.998	0.999	0.999	0.426	1.000	**0.597**	0.920	0.997	0.957
		0.998	0.998	**0.998**	0.907	0.997	0.950	0.555	1.000	**0.714**	0.998	0.998	0.998
		0.998	0.999	**0.998**	0.949	0.995	0.972	0.597	1.000	**0.748**	0.998	0.998	0.998
	2	0.999	0.996	**0.997**	0.935	0.994	0.963	0.324	1.000	**0.490**	0.789	0.997	0.881
		0.893	0.998	**0.943**	0.911	0.996	0.951	0.457	0.999	**0.627**	0.998	0.995	0.996
		0.912	0.990	0.949	0.917	0.978	0.946	0.615	0.959	**0.750**	0.998	0.996	0.997

Red bold values means indicates poor results, and the green bold values indicates ideally results.

signal photons but the misclassification of many noise photons as signal photons. This is because the Gaussian distribution difference of the photon density value between the signal points with high concentration and the remaining photon points is more significant than the difference between the signal and noise points. Similarly, the Gaussian distribution difference of the HDSTD value between the noise points with loose distribution in elevation and the remaining points is larger than the difference between the signal and noise points. The results of the two SGMs are closer to the real noise and signal distribution models by effectively considering the three statistics values simultaneously in multidimensional cases. It demonstrates the GMM method's ability to mitigate the respective weaknesses of different statistics and enhance the overall denoising effectiveness.

In low noise rate datasets, the denoising results for *k*-nearest neighbors and HDSTD are excellent, while the performance of photon density is subpar. This disparity is primarily due to the similarity of photon density values of noise and signal photons, misclassifying many signal photons as noise. However, incorporating photon density into the GMM method has a minimal negative impact, primarily due to its low differentiation in photon density and the GMM method's integration of two effective statistics, enhancing denoising quality.

Table III presents the four denoising results of the real ICESat-2 data. Nearly all datasets exhibit at least one negative statistic, while the GMM method still generates satisfactory results. The GMM method achieves classification with helpful statistics but is not susceptible to useless statistics. It sometimes selects valid data from negative statistics to attain superior results, reflecting its strong adaptability.

In daytime datasets, HDSTD consistently underperforms because the dataset distribution is not that clustered at elevation. In contrast, the GMM method enhances the strengths of *k*-nearest neighbors and photon density to yield impressive results. Even in some cases where both *k*-nearest neighbors and HDSTD have adverse effects, the GMM method closely approximates the best denoising results obtained with photon density alone. This is mainly due to the indistinguishable nature of negative statistics when incorporated as dimensions in a GMM, making the model approximately change from using 3-D statistics to 2-d statistics.

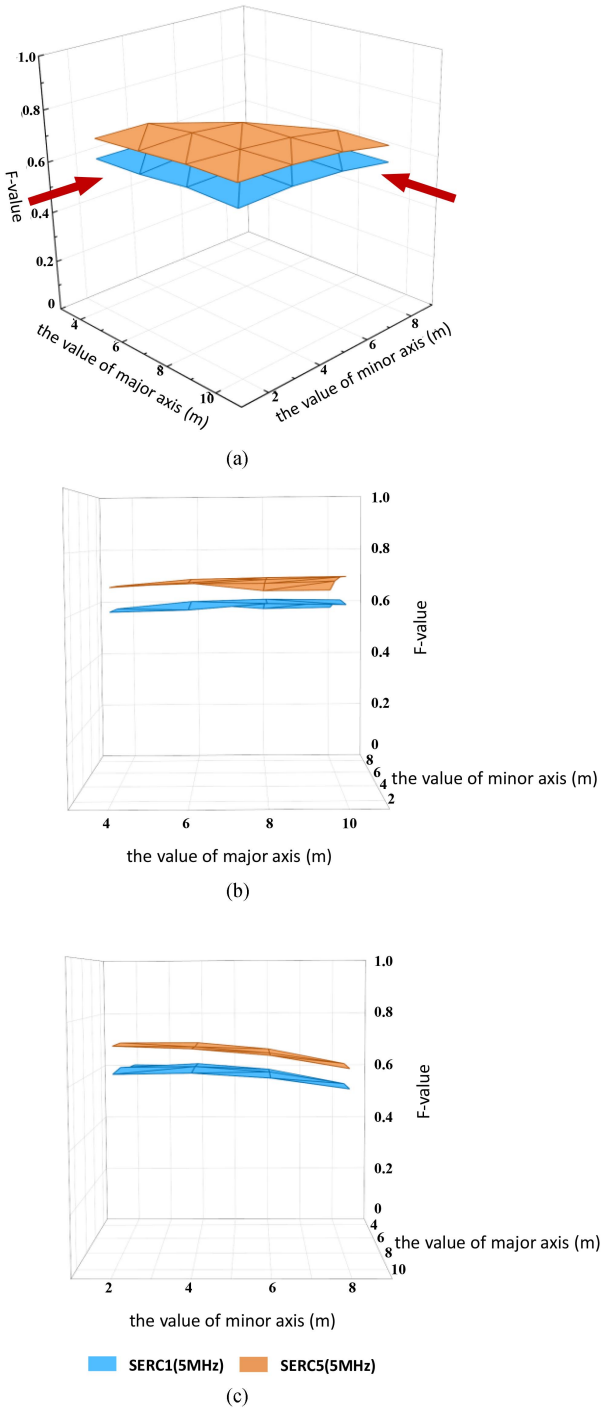


Fig. 23. Influence of the major and minor axes of the ellipse on the photon density. (a) The influence of ellipse axis on photon density. (b) The influence of major axis on photon density. (c) The influence of minor axis on photon density.

Thus, the negative statistics will contribute minimally to both classification and adverse effects.

In nighttime datasets, HDSTD performs well, while photon density underperforms. Similarly, considering the photon density's challenges, the GMM method consistently delivers denoising results that rival the best among the three statistics. Notably, the GMM method achieves the best results in the first nighttime dataset, even when a statistic has a detrimental effect. This indicates that although negative statistics are ineffective, combining them with other positive statistics within the GMM method can provide effective results.

Across all simulated and real datasets, the average *F*-value of the GMM-based multidimensional statistic is 84.73%, while the average *F*-values of the three single statistics, including the sum of the distance of *k*-nearest neighbors, photon density, and HDSTD, are 84.17%, 70.05%, and 80.98%, respectively. This demonstrates that the sum of the distances of *k*-nearest neighbors is the most useful statistic among the three statistics, and even if affected by the other two relatively poor statistics, the GMM method still obtains the best *F*-value among the four results. Since different statistics may appear to perform poorly in different datasets, the intelligent use of various statistics is truly significant. The GMM method achieves this function, making the denoising process easier and more accurate.

In order to comprehensively evaluate the denoising effect of the GMM method, two mainstream algorithms, including the DBSCAN and KNN methods, are applied for comparison, and the results demonstrate the excellent performance of the GMM method. To ensure comparability, the parameters to be set for the DBSCAN and KNN algorithms are consistent with those of the GMM algorithm. Specifically, for the DBSCAN algorithm, the neighborhood size, representing the major and minor axes of the ellipse, are set as 6 and 2 m, respectively. For the KNN algorithm, the number of nearest neighbors is set to 10.

Tables IV and V compare the final *F*-values of the simulated and real ICESat-2 data with respect to the DBSCAN and KNN methods. More intuitive differences are shown in Fig. 20 and Fig. 21. In simulated data, as the noise rate increases, the denoising results of the three algorithms decrease gradually. This indicates that the difference between signal and noise points is generally more remarkable in data with a low noise rate, making it easier to obtain better denoising results. Different denoising results can be obtained for various noise rates in real data, and the daytime datasets are inferior to the night datasets.

The GMM method always outperforms the DBSCAN and KNN methods for high and low noise rates. For all datasets, the average F-value of the GMM method is 85.46%, which is 5.22% and 7.26% higher than the corresponding ones for the DBSCAN and KNN methods (80.24% and 78.20%), respectively, reflecting the superiority of the GMM algorithm.

C. Parameter Sensitivity Analysis

The GMM-based multidimensional statistics denoising algorithm incorporates three types of statistics: the sum of distances of *k*-nearest neighbors, photon density, and HDSTD. This introduces parameters such as the *k*-value and the major and minor axes of the elliptical neighborhood. The SERC1 and SERC5 data with 5 MHz noise ratios of the simulated ICESat-2 dataset were utilized as examples to assess the sensitivity of these parameters.

Figs. 22 and 23 illustrate the influence of parameter variations on denoising results of the sum of distances of *k*-nearest neighbors and photon density. As shown in Fig. 22, the *F*-value is always stable as the *k* value increases. Fig. 23(a) shows the

TABLE IV
COMPARISON OF THE F-VALUES OF THE GMM, DBSCAN, AND KNN METHODS IN SIMULATED DATA

Datasets	Noise Rate	GMM			DBSCAN			KNN		
		R	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
Cedar2	0.5 MHz	0.949	0.981	**0.965**	0.895	0.971	0.932	0.826	0.975	0.895
	2 MHz	0.912	0.920	**0.916**	0.918	0.893	0.905	0.860	0.909	0.884
	5 MHz	0.842	0.766	**0.802**	0.959	0.559	0.706	0.929	0.612	0.738
Cedar4	0.5 MHz	0.953	0.983	**0.968**	0.888	0.972	0.928	0.816	0.976	0.889
	2 MHz	0.943	0.920	**0.931**	0.917	0.897	0.907	0.856	0.910	0.882
	5 MHz	0.892	0.734	**0.805**	0.966	0.597	0.738	0.935	0.645	0.764
SERC1	0.5 MHz	0.826	0.969	**0.892**	0.681	0.960	0.797	0.590	0.968	0.733
	2 MHz	0.866	0.811	**0.838**	0.729	0.844	0.782	0.647	0.863	0.739
	5 MHz	0.697	0.654	**0.675**	0.844	0.504	0.631	0.791	0.531	0.635
SERC3	0.5 MHz	0.869	0.977	**0.920**	0.746	0.971	0.844	0.652	0.980	0.783
	2 MHz	0.899	0.844	**0.871**	0.790	0.870	0.828	0.711	0.891	0.791
	5 MHz	0.744	0.671	**0.706**	0.889	0.530	0.664	0.843	0.576	0.685
SERC5	0.5 MHz	0.883	0.964	**0.922**	0.782	0.960	0.862	0.704	0.965	0.814
	2 MHz	0.907	0.858	**0.882**	0.815	0.872	0.842	0.747	0.890	0.812
	5 MHz	0.759	0.695	**0.726**	0.896	0.535	0.670	0.860	0.572	0.687

The green bold values indicates ideally results.

TABLE V
COMPARISON OF THE F-VALUES OF THE GMM, DBSCAN, AND KNN METHODS IN REAL DATA

Datasets	Number	GMM			DBSCAN			KNN		
		R	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
Daytime	1	0.889	0.946	**0.917**	0.728	0.992	0.837	0.737	0.998	0.848
		0.805	0.995	**0.890**	0.680	0.998	0.809	0.777	0.995	0.873
		0.923	0.993	**0.957**	0.882	0.993	0.933	0.971	0.879	0.923
	2	0.843	0.881	**0.860**	0.741	0.726	0.731	0.335	0.728	0.458
		0.872	0.763	**0.814**	0.757	0.586	0.656	0.741	0.770	0.755
		0.704	0.950	**0.809**	0.595	0.781	0.663	0.345	0.876	0.495
Nighttime	1	0.998	0.999	**0.999**	0.988	0.975	0.981	0.914	0.965	0.938
		0.998	0.998	**0.998**	0.988	0.974	0.980	0.909	0.962	0.935
		0.998	0.999	**0.998**	0.997	0.975	0.986	0.876	0.956	0.914
	2	0.999	0.996	**0.997**	0.941	0.999	0.969	0.928	0.957	0.942
		0.893	0.998	**0.943**	0.866	0.999	0.928	0.887	0.962	0.923
		0.912	0.997	**0.953**	0.854	1.000	0.921	0.901	0.965	0.932

The green bold values indicates ideally results.

combined effect of different values of the major and minor axes of an ellipse on the *F*-value. The major and minor axes of the ellipse correspond to the *x* and *y* coordinates of the 3-D graph, while the *F*-value corresponds to the *z* coordinate. Fig. 23(b) and (c) respectively provide observations of the three-dimensional graph from the directions of the major and minor axes of the ellipse [as indicated by the arrows in Fig. 23(a)] to better demonstrate the parameter insensitivity of the major and minor axes of the ellipse. The *F*-value changes little no matter how the values of the two axes change. To sum up, the GMM method can analyze the general characteristics of signal and noise points for denoising and is hardly affected by the overall change of the input value caused by the parameter adjustment, indicating its insensitivity to the input parameters.

V. CONCLUSION

This research proposed a multidimensional statistics denoising algorithm based on the GMM. The results were compared and analyzed comprehensively. The conclusions are as follows.

1) Compared multidimensional statistics with three single statistics, the former obtained the best results for effective inputs using multidimensional features to achieve complementarity.
2) Although some statistics misbehaved, the multidimensional statistics obtained relatively good results by eliminating adverse effects.
3) The GMM method does not require a threshold setting and can adaptively differentiate between signal and noise photons.
4) The GMM method based on multidimensional statistics outperformed the DBSCAN and KNN methods.
5) The GMM method is insensitive to the input parameters of the multidimensional statistics.

In summary, our proposed GMM method fully uses multidimensional statistics and is robust and adaptive for extracting signal photons. However, this study only selected several datasets from China, utilized the real ICESat-2 data collected in 2019 and 2020, and employed both the real and simulated data to perform the experiment. Future research should globally

select more datasets to evaluate the GMM method's reliability. As a future improvement, the noise reduction process should homogenize the noise level along the track direction to resolve the uneven distribution of noise densities.

ACKNOWLEDGMENT

The authors greatly appreciate the ICESat-2 Project Science Office for providing the simulated ICESat-2 data, and the National Snow and Ice Data Center for providing the ICESat-2 data. For inquiries about the code implementation of the GMM algorithm, contact the corresponding author.

REFERENCES

[1] X. Zhu et al., "Research progress of ICESat-2/ATLAS data processing and applications," *Infrared Laser Eng.*, vol. 49, no. 11, pp. 76–85, 2020.
[2] Y. Pang et al., "Status and development of forest carbon storage remote sensing satellites," *Spacecraft Recovery Remote Sens.*, vol. 43, no. 6, pp. 1–15, 2022.
[3] Y. Ma et al., "Satellite-derived bathymetry using the ICESat-2 lidar and Sentinel-2 imagery datasets," *Remote Sens. Environ.*, vol. 250, 2020, Art. no. 112047.
[4] X. Zhu, *Forest Height Retrieval of China with a Resolution of 30 m Using ICESat-2 and GEDI Data*, Aerospace Inf. Res. Inst., Chin. Acad. Sci., Beijing, China, 2021.
[5] B. Cao et al., "Point clouds denoising of photon counting LiDAR based on poisson distribution," *Hydrographic Surveying Charting*, vol. 42, no. 2, pp. 65–69, 2022.
[6] H. Li et al., "Analysis on ranging error of terrain targets for ICESat-2 single-photon lidar," *Infrared Laser Eng.*, vol. 49, no. 11, 2020, Art. no. 20200247.
[7] I. Ullah, *Deeply Supervised Approaches For Salient Object Detection in Complex Scenarios*, Shandong Univ., Jinan, China, 2021.
[8] B. Chen, "Photon counting lidar data processing and forest parameters estimation," Chin. Acad. Forestry, Beijing, China, 2019.
[9] D. Zhang, X. Lu, H. Qin, and Y. He, "Pointfilter: Point cloud filtering via encoder-decoder modeling," *IEEE Trans. Visual. Comput. Graph.*, vol. 27, no. 3, pp. 2015–2027, Mar. 2021.
[10] D. Lu, "Research On ICESat-2 photon point cloud denoising classification method," Kunming Univ. Sci. Technol., Kunming, China, 2022.
[11] X. Xu et al., "TDNet: Transformer-based network for point cloud denoising," *Appl. Opt.*, vol. 61, no. 6, pp. C80–C88, 2022.
[12] X. Guo, "A Study on image clustering algorithms with deep neural networks," Graduate School Nat. Univ. Defense Technol., Changsha, China, 2020.
[13] J. Zhang and J. Kerekes, "An adaptive density- based model for extracting surface returns from photon-counting laser altimeter data," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 726–730, Apr. 2015.
[14] X. Zhu et al., "A noise removal algorithm based on OPTICS for photon-counting LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1471–1475, Aug. 2021.
[15] S. Popescu et al., "Photon counting lidar: An adaptive ground and canopy height retrieval algorithm for ICESat-2 data," *Remote Sens. Environ.*, vol. 208, pp. 154–170, 2018.
[16] B. Chen et al., "Ground and top of canopy extraction from photon-counting LiDAR data using local outlier factor with ellipse searching area," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1447–1451, Sep. 2019.
[17] B. Chen and P. Yong, "A denoising approach for detection of canopy and ground from ICESat-2's airborne simulator data in Maryland, USA," *Appl. Opt. Photon. China*, 2015.
[18] L. A. Magruder et al., "Noise filtering techniques for photon-counting lidar data," *Laser Radar Technol. Appl. XVII*, vol. 8379, pp. 237–245, 2012.
[19] M. Awadallah et al., "Active contour models for extracting ground and forest canopy curves from discrete laser altimeter data," in *Proc. 13th Int. Conf. LiDAR Appl. Assessing Forest Ecosyst.*, 2013, pp. 129–136.
[20] Z. Hui et al., "A novel denoising algorithm for airborne LiDAR point cloud based on empirical mode decomposition," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 1021–1025, 2019.
[21] H. Tang et al., "Voxel-based spatial filtering method for canopy height retrieval from airborne single-photon lidar," *Remote Sens.*, vol. 8, no. 9, 2016, Art. no. 771.
[22] M. Li et al., "A noise filter method for the Push broom photon counting lidar and airborne cloud data verification," *Sci. Technol. Eng.*, vol. 17, no. 9, pp. 53–58, 2017.
[23] S. Xia et al., "Point cloud filtering and tree height estimation using airborne experiment data of ICESat-2," *J. Remote Sens.*, vol. 18, no. 4, pp. 1199–1207, 2014.
[24] S. Nie et al., "Estimating the vegetation canopy height using micro-pulse photon-counting LiDAR data," *Opt. Exp.*, vol. 26, no. 10, pp. A520–A540, 2018.
[25] D. Gwenzi et al., "Prospects of the ICESat-2 laser altimetry mission for savanna ecosystem structural studies based on airborne simulation data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 118, pp. 68–82, 2016.
[26] G. Yang, S. Rong, and M. Li, "An adaptive directional filter for photon counting lidar point cloud data," *J. Infrared Millimeter Waves*, vol. 36, no. 1, pp. 107–113, 2017.
[27] C. Pan et al., "Single photon point cloud denoising method based on density and local statistics," *J. Univ. Chin. Acad. Sci.*, vol. 80, 2022, pp. 268–274.
[28] A. Neuenschwander and L. Magruder, "Canopy and terrain height retrievals with ICESat-2: A first look," *Remote Sens.*, vol. 11, no. 14, 2019, Art. no. 1721.
[29] X. Wang, Z. Pan, and C. Glennie, "A novel noise filtering model for photon-counting laser altimeter data," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 7, pp. 947–951, Jul. 2016.
[30] 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 Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 6501405.
[31] X. Zhu et al., "A ground elevation and vegetation height retrieval algorithm using micro-pulse photon-counting lidar data," *Remote Sens.*, no. 10, 2018, Art. no. 1962.
[32] H. Wang, "Research on hand gesture tracking based on adaptive active contour model," in *Lanzhou Univ. Technol.*, Lanzhou, China, 2012.
[33] Q. Wang and D. Hu, "Optimization of K-means fast clustering algorithm based on Spark," *Comput. Simul.*, vol. 39, no. 3, pp. 344–349, 2022.
[34] T. Lin and C. Zhao, "Nearest neighbor optimization k-means clustering algorithm," *Comput. Sci.*, vol. 46, no. S2, pp. 216–219, 2019.
[35] S. Xia et al., "Point cloud filtering and tree height estimation using airborne experiment data of ICESat-2," *J. Remote Sens.*, vol. 18, no. 6, pp. 1199–1207, 2014.
[36] H. Xie, Q. Xu, and D. Ye, "A comparison and review of surface detection methods using MBL, MABEL, and ICESat-2 photon-counting laser altimetry data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7604–7623, Jul. 2021.
[37] J. Liu and J. Liu, "Analysis and comparison of denoising methods for photon-counting laser data," in *Proc. 3rd Int. Conf. Geol., Mapping Remote Sens.*, 2022, pp. 175–178.
[38] B. Chen et al., "Photon-counting LiDAR point cloud data filtering based on the random forest algorithm," *J. Geo-Inf. Sci.*, vol. 21, no. 6, pp. 898–906, 2019.
[39] D. Lu et al., "Denoising and classification of ICESat-2 photon point cloud based on convolutional neural network," *J. Geo-Inf. Sci.*, vol. 23, no. 11, pp. 2086–2095, 2021.
[40] S. Wei et al., "A single photon denoising algorithm combined with improved DBSCAN and statistical filtering," *Laser Technol.*, vol. 45, pp. 601–606, 2020.
[41] G. Zhang, S. Xing, Q. Xu, F. Zhang, M. Dai, and D. Wang, "An automatic algorithm to extract nearshore bathymetric photons using pre-pruning quadtree isolation for ICESat-2 data," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Mar. 2023, Art. no. 1501505.
[42] S. Wei et al., "Single-photon denoising algorithm based on line scanning characteristics of lidar channels," *Laser Optoelectron. Prog.*, vol. 59, no. 12, pp. 116–122, 2022.
[43] C. Wang et al., "Adaptive denoising algorithm for photon-counting LiDAR point clouds," *Laser Optoelectron. Prog.*, vol. 58, no. 14, pp. 482–489, 2021.
[44] L. Qin et al., "Adaptive denoising and classification algorithms for ICESat-2 airborne experimental photon cloud data of 2018," *J. Remote Sens. (Chin.)*, vol. 24, no. 12, pp. 1476–1487, 2020.
[45] P. Yang, H. Fu, J. Zhu, Y. Li, and C. Wang, "An elliptical distance based photon point cloud filtering method in forest area," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 2022, Art. no. 6504705.
[46] Y. Pang et al., "Status and development of spaceborne lidar applications in forestry," *Aerosp. Shanghai*, vol. 36, no. 3, pp. 20–27, 2019.

[47] W. Liu, T. Jin, J. Li, and W. Jiang, "Adaptive clustering-based method for ICESat-2 Sea Ice retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4301814.
[48] J. Shan et al., "Advances of spaceborne laser altimetry technology," *Acta Geodaetica et Cartographica Sin.*, vol. 51, no. 6, pp. 964–982, 2022.
[49] Z. Wang, X. Xi, S. Nie, and C. Wang, "Bathymetric method of nearshore based on ICESat-2/ATLAS data—A case study of the islands and reefs in The South China Sea," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 2868–2871.
[50] X. Pu and H. Xiao, "Application of recurrence plot analysis method to fault diagnosis based on EMD," *Mach. Tool Hydraul.*, vol. 43, no. 5, pp. 160–163, 2015.
[51] Y. Li, H. Fu, J. Zhu, and C. Wang, "A filtering method for ICESat-2 photon point cloud data based on relative neighboring relationship and local weighted distance statistics," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 1891–1895, Nov. 2021.
[52] S. Si et al., "Multiscale feature fusion for the multistage denoising of airborne single photon LiDAR," *Remote Sens*, vol. 15, no. 1, 2023, Art. no. 269.
[53] A. Dittrich, M. Weinmann, and S. Hinz, "Analytical and numerical investigations on the accuracy and robustness of geometric features extracted from 3D point cloud data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 126, pp. 195–208, 2017.
[54] B. Cao et al., "ICESAT-2 shallow bathymetric mapping based on a size and direction adaptive filtering algorithm," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6279–6295, Jun. 2023.
[55] M. Vicari et al., "Leaf and wood classification framework for terrestrial LiDAR point clouds," *Methods Ecol. Evol.*, vol. 10, pp. 680–694, 2019.
[56] X. Tian and J. Shan, "Detection of signal and ground photons from ICESat-2 ATL03 data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Dec. 2023, Art. no. 5700314.
[57] L. Ma et al., "Determining woody-to-total area ratio using terrestrial laser scanning (TLS)," *Agricultural Forest Meteorol.*, vol. 228, pp. 217–228, 2016.
[58] B. Zhang et al., "Improved forest signal detection for space-borne photon-counting LiDAR using automatic machine learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1–13, Jun. 2024.
[59] X. Zheng, C. Hou, M. Huang, D. Ma, and M. Li, "A density and distance-based method for ICESat-2 photon-counting data denoising," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Feb. 2023, Art. no. 6500405.

Sitong Chen received the B.E. degree in navigation engineering from Wuhan University, Wuhan, China, in 2023. She is currently working toward the M.S. degree in cartography and geographic information system from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

Her research interests include data denoising and classification of photon-counting LiDAR data.

Sheng Nie received the Ph.D. degree in cartography and geographic information system from the University of Chinese Academy of Sciences, Beijing, China, in 2017.

He is an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include light detection and ranging data processing and its application in forestry.

Xiaohuan Xi received the M.S. degree in environmental geography from Peking University, Beijing, China, in 2000.

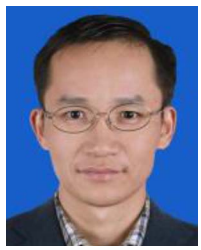
She is an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. Her research interests include forest characteristics retrievals using LiDAR data.

Shaobo Xia received the Ph.D. degree in geomatics from the University of Calgary, Calgary, AB, Canada, in 2020.

He is an Assistant Professor with the Department of Geomatics Engineering, Changsha University of Science and Technology, Changsha, China. His research interests include point cloud processing, image processing, and remote sensing.

Feng Zhu received the B.Sc., master's, and Ph.D. degrees in geodesy and engineering surveying with distinction in geodesy and engineering surveying from the School of Geodesy and Geomatics, Wuhan University, Wuhan, China, in 2012, 2015, and 2019, respectively.

He is an Associate Professor with the School of Geodesy and Geomatics, Wuhan University. His current research interests include multisensor integration and its application in position and orientation systems and autonomous systems.



Cheng Wang received the Ph.D. degree in remote sensing from Université Louis Pasteur, Strasbourg, France, in 2005.

He is a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, and also with the College of Resources of Environment, University of the Chinese Academy of Sciences, Beijing. His research interests include LiDAR remote sensing mechanisms and multimode LiDAR data processing.

Xiaoxiao Zhu received the Ph.D. degree in cartography and geographic information system from the University of Chinese Academy of Sciences, Beijing, China, in 2021.

She is an Associate Professor with the Institute of Geology, China Earthquake Administration, Beijing. Her research interests include data processing and application of spaceborne LiDAR.