NSANet: Noise-Seeking Attention Network

Maryam Jameela [iD] and Gunho Sohn [iD], *Member, IEEE*

Abstract—Light detection and ranging technology has remained popular in capturing natural and built environments for numerous applications. The recent technological advancements in electrooptical engineering have aided in obtaining laser returns at a higher pulse-repetition frequency (PRF), which considerably increased the density of the 3-D point cloud. Conventional techniques with lower PRF had a single pulse-in-air zone, large enough to avoid a mismatch among pulse pairs at the receiver. New multiple-pulse-in-air technology guarantees various windows of operational ranges for a single flight line and no blind zones. The disadvantage of the technology is the projection of atmospheric returns closer to the same pulse-in-air zone of adjacent terrain points likely to intersect with objects of interest. These noise properties compromise the perceived quality of the scene and encourage the development of new noise-filtering neural networks, as the existing filters are significantly ineffective. We propose a novel dual-attention noise-filtering neural network called noise-seeking attention network (NSANet) that fuses physical priors and local spatial attention to filter noise. Our research's fusion module is motivated by two psychological theories of feature integration and attention engagement to prove the role of attention in computer vision at the encoding and decoding phase. The presented results of NSANet show the benefit of attentional engagement theory and a performance boost of 7.30% on recall and 4.10% on F1-score compared to the state-of-the-art noise-filtering deep convolutional neural networks.

Index Terms—Attention, computer vision, noise-filtering neural network, physical priors, point cloud, systematic noise.

I. INTRODUCTION

LIGHT detection and ranging (LiDAR) is a critical tool for mapping natural and man-made environments in military and civilian applications. Recent advancements in electrooptical engineering have allowed for high pulse-repetition frequency (PRF) laser returns, leading to significantly denser 3-D point clouds. Airborne LiDARs, also known as active sensors, emit pulses and generate reflections from backscattered signals.

Compared to other data acquisition methods, airborne LiDAR surveys are faster and more accurate due to their maturity and the following characteristics: 1) high-density point cloud capture ability; 2) recording of discrete or full-waveform backscattered signals, allowing for mapping of ground surfaces, low vegetation, and even forest and heavily populated urban areas;

3) no blind spots in the sensor technology; 4) faster surveys leading to economical data acquisition; and 5) reduced impact from extreme weather or environmental conditions on resulting mapping .

Despite the capabilities of LiDAR to capture the reality of a scene, the quality can be compromised by unwanted atmospheric data, referred to as noise [1]. Laser scanner technology has come a long way since its early days when it could only manage a few thousand pulses per second (pps) and could only capture the first and last pulse return with no information on intensity. Today, thanks to technological progress, laser scanners can manage up to half a billion pps, allowing for the capture of both full-waveform and discrete data while accurately measuring the energy of the reflected signal.

The main goal of developing LiDAR technology has been to achieve high point density with a relatively smaller mean point spacing. This objective was driven by the need for fast, efficient, and detailed data collection, which required a higher PRF, greater altitude, and an increased scan field of view (SFOV) for wider coverage. The previous single-pulse-in-air (SPIA) technology did not meet these requirements [1].

In airborne LiDAR systems, the term "pulse-in-air (PIA) zone" refers to a maximum unambiguous range, where the laser pulse propagates through the air before reaching the ground or target surface, and the next beam is yet to be fired, as shown in Fig. 3.

The conventional LiDAR technology performs the survey using an SPIA zone with a lower PRF. In this system, a single pulse is fired, and the transmitter waits for a time interval of 1/PRF before shooting the next pulse and receiving the previous signal before it. Fig. 2(a) illustrates that at a certain time, only one pulse is in the air, such as S_2 is not fired before R_1 is received. These systems are slow, costly, and have only one operational range per flight line. Moreover, they generate less detailed data and have a relatively lower point density [2].

To overcome the limitations of conventional LiDAR technology, multiple-pulse-in-air (MPIA) technology was introduced in 2006. Unlike the SPIA system, MPIA fires multiple pulses without waiting for them to return. The receiver then matches the returned signals with all the pulses fired using proprietary dithering algorithms [3]. MPIA technology is highly configurable, allowing intensity and PRF adjustments to acquire more detailed high-density point clouds. This results in fast cost-effective gate-less sensors that do not generate blind zones. These sensors provide multiple operational ranges for a single flight. As illustrated in Fig. 2(b), at a certain time, there could be multiple pulses in the air, such as S_1 and S_2, before R_1 is reflected.
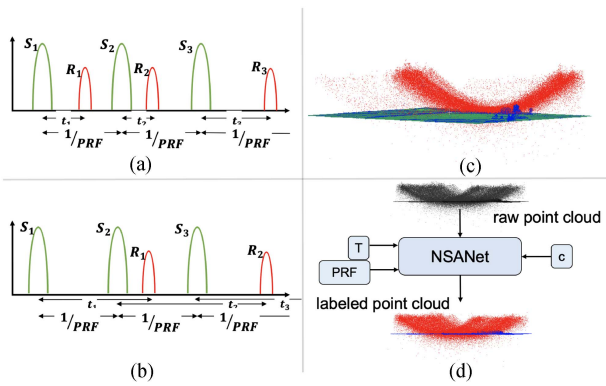
Fig. 1. (a) SPIA is a traditional technology, which matched the signal sent by a transmitter (S) and received by a receiver (R) efficiently but was expensive. (b) MPIA technology improved the cost and speed up the data collection though it is more prone to atmospheric points, which can be seen overlapping with objects of interest in (c). (d) Proposed NSANet uses physical priors and spatial attention to filter noise from the point cloud.
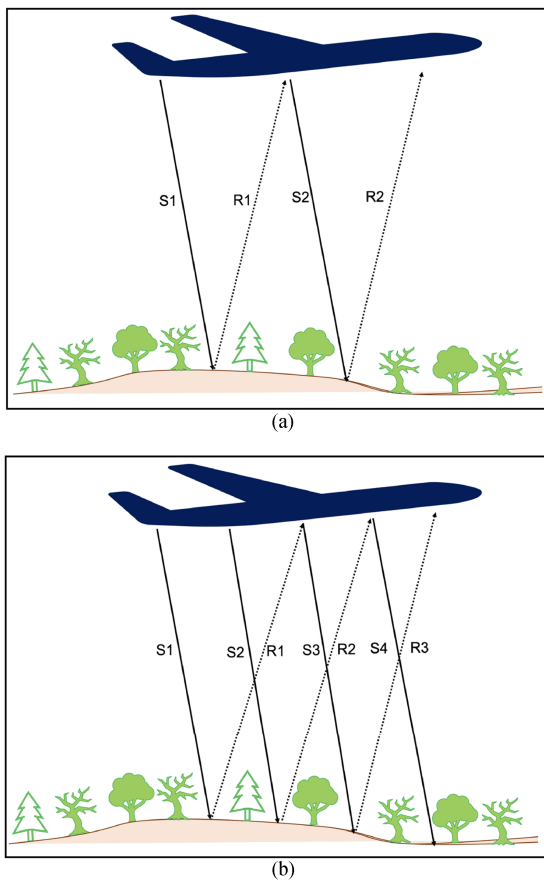

Fig. 2. (a) SPIA zone shows one pulse sent and received before next is fired. (b) MPIA fires second pulse before first is returned.

The use of MPIA technology in LiDAR systems can be affected by atmospheric points caused by environmental factors, such as dust, fog, snow, or rain particles. When the laser beam is deflected or scattered by these atmospheric particles, it can result in the misinterpretation of signals on the receiver as objects of interest, generating noise [4].

Increasing the PRF has a detrimental effect as it narrows the PIA zones and increases the difficulty of differentiating noise points from objects of interest. Consequently, the projection of these noise points can give the impression that they are closer to the objects of interest and occur during the transition among PIA zones, as depicted in Fig. 3. This proximity makes distinguishing the noise points from the objects of interest difficult.

Resolving range ambiguity is crucial in the postprocessing analysis of acquired data. Range ambiguity arises when the returned pulses overlap on the receiver. The noise points are projected onto the ground at the closest possible range to address range ambiguity. This is necessary because noise points detected at longer ranges can falsely appear closer to the ground.

Consequently, when the noise points are projected, they can give the illusion of being closer to the transition zones among different PIA areas. This phenomenon occurs because the range ambiguity resolution process places the noise points at ranges closer to the ground, creating the perception that they are adjacent to the transition zones. Fig. 1(c) depicts the compromised quality of perception due to noise [5].

Our data show variable noise density exhibiting characteristics, such as complex objects (i.e., trees and bushes). Furthermore, the global systematic noise pattern is localized on the transition point between two PIA zones and overlaps with objects of interest. Another major characteristic is imbalanced class distribution, i.e., noise points making up (1–3)% of the complete scene.

While the atmospheric returns always subsisted at lower PRF, we could eliminate them with simple height filtering or nearest neighboring algorithms [6]. However, more sophisticated filtering algorithms are required once noise begins to overlap with the key objects and compromise the scene's semantics.

Deep convolutional neural networks (CNNs) have surpassed traditional computer vision techniques in recent times [7], [8], [9], [10]. They can emulate human vision for perceiving 3-D scenes by extracting complex and deep features to learn patterns for detection, classification, and semantic segmentation. This inspired us to develop a large-scale noise classification deep neural network that efficiently and robustly filters noise from the point cloud. One of the key limitations in the existing noise-filtering literature is that the proposed techniques have not been used to remove noise from large-scale 3-D scenes [11], [12], [13], [14], [15], [16], [17]. The remote sensing community mostly treats noise filtering as an outlier detection problem for noise in 3-D objects. This neglects the problem most manufacturers face to remove the noise added in the dithering step due to incorrect projections of atmospheric points closer to objects of interest.

3-D large-scale point cloud noise-filtering techniques were not developed largely by the research community due to the manufacturers' in-house techniques to filter the noise and release a relatively cleaner dataset for commercial and academic use. These techniques were based on height filters and the support vector machine (SVM), and now, Teledyne Optech is using the method proposed in this article. This is a preprocessing technique. Therefore, the major concern for the selection is a good tradeoff between the speed of inference and the effectiveness of the network. Our system design was motivated by various

noise types and physical priors branching into physics-based deep learning.

As mentioned earlier, noise can be classified into two types based on localization: local and global. To address this, we propose a system that fuses these spatial contexts using attention inspired by human attention phenomena. The basic idea of attention in computer vision is to withdraw attention from irrelevant objects and focus on relevant objects by assigning more weight to relevant features. Different types of attention implementation exist, such as additive, multiplicative, self-attention, and others. Our study focuses on using additive and multiplicative attention.

Psychologists have studied the fusion of attention into the decision-making process and proposed two relevant theories: feature integration theory (FIT) proposed by Treisman and Gelade [18] and attentional engagement theory (AET) proposed by Duncan and Humphreys [19]. FIT backs the fusion of attention at the encoding and decoding stages, while AET does it only at the decoding stage. We designed our network using both the fusion theories.

Our proposed method, the noise-seeking attention network (NSANet), is a novel prior-based dual-attention noise-filtering neural network. Our key contributions are as follows.

1) We developed a dual-attention-based noise-filtering neural network for the large-scale 3-D point cloud.
2) We proposed a dual-attention module, which embeds physical priors and local spatial attention to guide the network to activate a global noise pattern and eliminate ambiguity among overlapping objects.
3) We designed an ablation study focused on the fusion of attention into a network based on two psychological theories: FIT and AET.

The following sections discuss related work, motivation, methodology, experimental configuration, and results.

II. LITERATURE REVIEW

This section will be divided into four parts: 1) introduction to airborne laser scanning; 2) physical-prior-based network; 3) noise-filtering networks; and 4) multiscale feature fusion.

A. Airborne Laser Scanning: PIA

Airborne LiDARs are active sensors emitting pulses and generating backscattered signal reflections. Over the past few decades, remote sensing technology has progressed, and airborne LiDAR surveys have become more common. They are used to plan for natural and human disasters, which can be prevented with the aid of this technology. In contrast to other data acquisition methods and tools, airborne LiDARs are more efficient, faster, and accurate due to the technology's maturity and the following vital characteristics: 1) their ability to capture an exact and highly dense point cloud; 2) LiDARs can record discrete or full waveform of the backscattered signal, map ground surfaces, and detect low vegetation even in forests and heavily populated urban areas; 3) the sensors have no blind zones; 4) surveys are faster, making them an economical data acquisition method; and 5) extreme weather or environmental conditions have a relatively lower effect on the resultant mapping [3].

Point clouds' digitization using full or discrete waveform has advantages and disadvantages. Full-waveform digitization captures vertical structures in a time-series dataset, while discrete waveform only captures peaks. Since the thesis only utilizes discrete-waveform-based LiDAR, this section will focus solely on airborne LiDARs based on discrete waveform technology [3].

The airborne LiDAR system is a complex multisensor device with various essential components, such as a reflectorless laser sensor, an airborne inertial measuring unit (AIMU), a global positioning system (GPS), data storage, and a power supply. As the system carrier, such as an aircraft or plane, flies over the area of interest, the laser sensor emits pulses toward the surface. These pulses travel at the speed of light and reflect to the sensor, depending on the surface's reflectivity. This enables the system to create a highly accurate 3-D map of the area of interest, which can be used for various applications in fields, such as surveying, mapping, and environmental management.

After receiving the reflected energy, the sensor calculates a 3-D point based on the distance traveled by the pulse, the sensor's position, and the beam's direction. To track the LiDAR's position accurately, the system uses trajectory information from a global navigation satellite system receiver, which provides altitude and orientation. In addition, the AIMU records the LiDAR's position using pitch, roll, and yaw angles. The LiDAR's signal is deflected using an internal mirror, and the mirror's position is recorded for every laser pulse shot, as shown in Fig. 4. Finally, the range between the target and the sensor is calculated using the Euclidean distance as follows:

$$R = \sqrt{(x_p - x_l)^2 + (y_p - y_l)^2 + (z_p - z_l)^2} \qquad (1)$$

where x_p, y_p, and z_p are the coordinates of the points and x_l, y_l, and z_l are the coordinates of the sensor.

Advancements in LiDAR technology have enabled a high PRF of over half a million kilohertz. The PRF represents the ability to fire the pulses within 1 s, calculating how many pulses can travel with the speed of light (c) to the maximum unambiguous range R_{max} in 1 s, as follows:

$$PRF = \frac{c}{2 \times R_{max}}. \qquad (2)$$

This enhances survey efficiency, allowing data acquisition with fewer flight lines and a wider SFOV. The SFOV is the angle at which LiDAR emits the pulses, set based on the target application and sensor range. Different acquisition techniques have been used for effective, efficient, and fast surveys. The PIA serves as the basis for acquisition technology, allowing the measurement of how many pulses can be in the air at a given time or before the next one is fired. The PIA zone is the space between the laser and the target, where a beam travels through before the next pulse is fired. The number of total PIA zones depends on the following equation:

$$PIA = \frac{PRF/(c/2)}{Altitude/\cos(SFOV)}. \qquad (3)$$

SPIA is a traditional technology that was discarded due to its inability to match the high PRF. MPIA technology has been
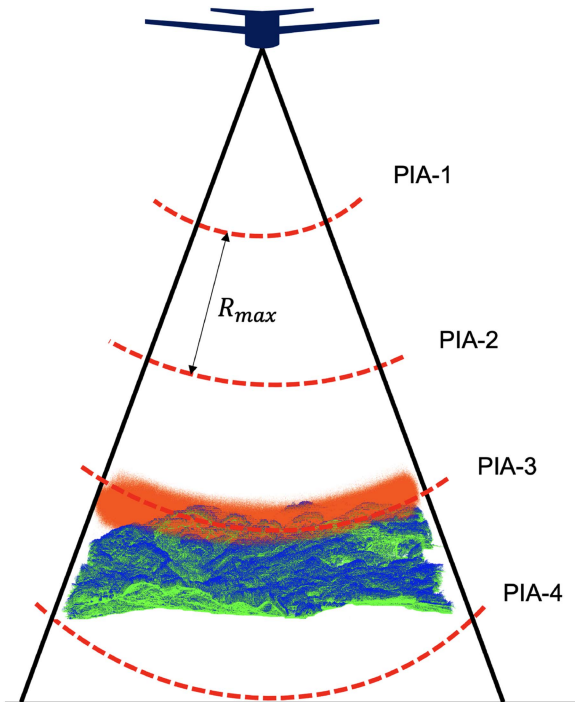
Fig. 3. PIA zones and projection of objects and noise points on transition. R_{max} shows the range a pulse travels before the next one is fired.
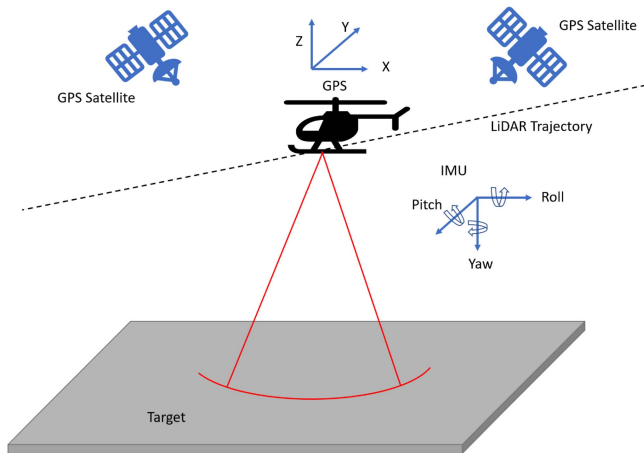

Fig. 4. Airborne LiDAR.

developed, surpassing the limitations of traditional SPIA technology [4]. This technology ensures the firing of multiple pulses before receiving the last one, making it efficient and fast, as depicted in Fig. 3. This thesis largely focuses on the 3-D point cloud acquired using MPIA technology for noise filtering and utility semantic analysis.

1) Single Pulse in Air: SPIA is a conventional method for data acquisition in airborne LiDAR, which is cost effective at low PRF. SPIA involves discrete data acquisition, firing the next pulse only when the previous one is received. However, at higher PRF, SPIA technology struggles to keep up due to its limitation of allowing only one pulse in the air at a time. This drawback can lead to missing or corrupted data points, especially at higher

altitudes or higher PRF speeds where the speed of light is not fast enough to return the pulse before the next one is fired [20].

2) Multiple Pulses in Air: In contrast to SPIA technology, MPIA technology operates more efficiently by not waiting for a pulse to be fully reflected and matched before firing the next one, as illustrated in Fig. 2. However, this efficiency has a drawback— the laser pulse pairs across PIA zones can result in atmospheric points being projected closer to or overlapping with objects of interest. Consequently, the point cloud obtained through MPIA technology may suffer from compromised scene perception due to noise.

In the past, when atmospheric returns were present at lower PRF, they could be easily removed using simple height filters or nearest neighbor algorithms. However, with MPIA, these atmospheric points started to mix with the objects of interest, necessitating more sophisticated algorithms for noise filtering. In addition, this technology affects the projection of wire points in the correct PIA zone during range resolution for the digitization of the 3-D point cloud. These issues have encouraged the development of dithering algorithms. Various companies have employed proprietary algorithms, which are not disclosed [21]. Hence, the thesis focuses on cleaning the 3-D point cloud obtained through MPIA technology, such as Teledyne Optech Galaxy T1000, and using it to segment the utility scenes correctly.

It is important to understand that airborne LiDAR can capture high-resolution dense 3-D point clouds for survey-grade applications [22].

B. Physics-Based Learning

Deep learning has significantly improved classification, object detection, recognition, and segmentation. However, integrating physical priors and principles into deep learning models, known as physics-based learning, has shown great potential for solving some problems [23]. P^3VAE is a generative model integrating deep learning with physics modeling to learn a high-dimensional latent space. It has shown improvements for semantic segmentation by regularizing the feature space [24]. A self-supervised dehazing algorithm also used physical priors to generate a lightweight network that outperformed the existing systems [25]. Despite various physics-based learning models that target complex scientific problems, there is a research gap in utilizing physical priors for semantic segmentation and denoising of 3-D point clouds [26].

C. Noise Filtering

The techniques for noise filtering can be categorized into four main groups: 1) statistical and geometrical methods; 2) clustering-based approaches; 3) machine learning techniques; and 4) deep learning algorithms.

1) Statistical and Geometrical Noise Filtering: The statistical filtering techniques used for removing noise from 3-D point clouds and images include univariate location and scale estimation, low-pass filters, mean filters, median filters, principal component analysis, multivariate location, covariance estimation, and spatial filters [27], [28], [29], [30], [31], [32], [33].

While these filters have been useful for cleaning data collected using the SPIA technology, they are ineffective and expensive for large-scale 3-D point clouds with high noise density and complexity.

Geometrical techniques have traditionally been used for outlier detection in 3-D point clouds, such as density [34], distance [35], graph [36], [37], [38], [39], and probability-based methods. These techniques were mostly employed for denoising 3-D meshes, models, and sparse point clouds. However, they have also been transformed into voxelization pipelines, and density- or distance-based filters have been applied to remove noise [40]. Bipartite graphs for approximating errors (noise) and manifold-to-manifold distance-based methods for outlier detection have shown good performance [41].

Nonetheless, geometrical methods are ineffective and expensive for large-scale 3-D point clouds with clustered and complex noise. For instance, the research work in [42] deals with a large-scale point cloud that contains single-echo reflection values. It has dealt with multisensor-based point clouds to remove reflection-based noise. However, it does not work for MPIA-based LiDAR due to multiple returns.

2) Cluster-Based Noise Filtering: It includes methods such as the nearest neighbor, radius outlier removal, mean-shift clustering, K-means, and DBScan [27], [43], [44], [45]. However, these methods require significant memory to construct clusters that adequately preserve object points while removing outliers and corrupted points. These work well for datasets with low overlap with objects of interest but can be resource intensive when covering each point and performing recursive clustering. Furthermore, these techniques are commonly used for 3-D point cloud models rather than large-scale scenes. While some modifications have been made to run these techniques on GPUs for a performance boost, they may still not be suitable for handling large-scale scenes effectively.

3) Machine-Learning-Based Noise Filtering: The emergence of machine learning has brought a significant impact on data analysis techniques, including the cleaning of 3-D point clouds. Linear regression, SVM, random forest, and other methods have been employed by various communities to filter or eliminate noise [46], [47]. These techniques have proven successful in preserving the object points while removing obvious noise. However, the tradeoff between maintaining scene quality and removing noise depends on the application domain's analysis. One of the challenges of using these techniques is the need for feature crafting and selection, which requires a thorough understanding of the dataset and domain. In recent years, researchers in noise filtering have explored the use of CNNs after introducing deep learning for scene understanding.

4) Deep-Learning-Based Noise Filtering: Deep learning in computer vision has significantly improved segmentation, classification, reconstruction, and detection. CNNs extract features from input data and detect decision-making features. They have shown performance boosts in denoising, surface reconstruction, densification, noise classification, and filtering. PointCleanNet [5], PointFilter [48], neural project denoising [49], MaskNet [50], and CNN-based normal estimation [51] for smoothing have improved filtering noise using various techniques

through deep neural networks. Recently, the transformer-based encoder–decoder model has been used to filter noise within point clouds containing single-echo reflection values [52]. It converts the point cloud into range images and removes noise. The work is efficient for single-echo reflective sensors but could not generate range images for airborne LiDAR using MPIA. Research is promising and shows potential for noise-filtering 3-D point clouds. The limitation of the previous research work is the lack of a dedicated network for a large-scale 3-D point cloud.

The existing deep learning methods have demonstrated the capabilities to remove noise from 3-D objects. The reason is that their generalization over 3-D large-scale scenes will not be effective and smooth object segmentation complexities, which in a large-scale scene can affect the performance of the denoising network. Most of the proposed networks are based on PointNet, and they have difficulty generalizing for denoising in large-scale 3-D scenes [17], [53]. They also have a scale sensitivity issue, which is hard to handle for large-scale noises in complex scenes as these networks are not built to cater to the global semantic context, which is key for filtering noise in a 3-D scene. These methods have also not shown comparable performance on complex and clamped noises.

One of the key aspects is to fuse the global and local contexts into the network for filtering noise. Our method considers multiscale feature fusion and attention to explore the scale sensitivity for filtering noise better. The next section will discuss attention and its use in computer vision.

D. Multiscale Feature Fusion

Feature fusion typically serves as a central module in feature aggregation. These modules necessitate embedding and fusing features from different scales to integrate context seamlessly into the deep learning network. Rather than simply concatenating features from different scales, these fusion modules capitalize on attention [54]. This approach enables the selecting of significant features for enhanced highlighting and performance improvement in the given task. To gain a deeper insight into feature fusion using attention, the following section outlines the attention mechanisms and psychological theories behind fusing features more effectively incorporating context.

1) Attention: This research explored using attention to embed physical priors into a neural network. Attention is a cognitive phenomenon that allows humans to focus on certain objects while ignoring others selectively. In deep learning, attention is a two-stage process, where the network first extracts features from the entire input and then emphasizes important details to improve the accuracy of tasks, such as classification, object detection, segmentation, and tracking [55]. Attention has been widely studied and adapted from human vision into computer vision. In fact, "Attention Is All You Need" [56] was the first publication to integrate attention mechanisms into natural language processing and sparked the use of attention in other deep learning tasks.

Over time, researchers have developed and studied various types of attention mechanisms, such as single head, multihead, additive, multiplicative, spatial, channel, and fusion attention [55]—the questions of where and what to attend to

have been major research areas in this field. For example, spatial self-attention has improved performance in segmentation and object detection [10]. In addition, additive self-attention has been found to work well for ambiguous boundaries among objects of interest. Psychological theories of attention guide this thesis. It utilizes a three-stage paradigm to embed physical priors and domain-based spatial layouts into neural networks for filtering noise and understanding large-scale 3-D point clouds:

1) a context-independent extraction of low-level features;
2) modeling relationship and analysis of low-level features are based on the computer vision task;
3) focus on relevant regions based on earlier stages of extracting high-level features for the final task [57].

Research focuses on understanding this fusion of attentional theories to propose the most appropriate embedding technique of physical priors into the deep learning model.

2) Attentional Engagement Theory: AET, as introduced by Duncan and Humphreys [19], outlined the attention process into two distinct stages: 1) the preattentive stage, where features are extracted without focal attention, and 2) the selective attention stage, where these features are intensely and focused upon to facilitate decision making.

In the realm of computer vision and deep learning, this theory can be translated into a two-part model, where, initially, an attention-free encoder processes features from the input data. This corresponds to the preattentive stage of AET, where features are detected without attentional focus. Such an encoder would scan the entire dataset, identifying potential interest features without determining their significance or role.

Subsequently, the selective attention stage of AET is mirrored in the form of an attention-based decoder. This model component is designed to selectively concentrate on the features encoded by the first stage, applying intense focus to discern patterns, make connections, and ultimately assist in decision making. This could manifest as a network that, having encoded various features, now applies a "filter" to decide which features are most pertinent to the task, such as segmenting specific structures from a medical image or identifying objects within a complex scene.

The principles of AET are exemplified in deep learning architectures, such as the Attention 3-D U-Net [10]. These networks leverage attention mechanisms to enhance the feature selection process during the decoding phase, ensuring that the network pays "attention" to the most informative parts of the data for accurate segmentation or detection.

Incorporating AET into deep learning models provides a framework that can handle complex noisy datasets by mimicking the human attention process—initially surveying and extracting a broad array of information and then honing in on the most critical features for the task. This allows for more precise and relevant feature utilization in object detection and segmentation tasks.

3) Feature Integration Theory: According to FIT, proposed by Treisman and Gelade [18], the visual system processes individual features, such as color, shape, and size in a parallel and preattentive manner. However, integrating these features into a coherent object perception requires focused attention.

In the scope of this thesis, the theory is operationalized through the design of an attention-based encoder and decoder framework. The encoder emulates the preattentive stage, capturing and encoding disparate features independently. These features are processed in parallel, akin to the initial stage of visual processing, where the brain registers various attributes of objects.

Conversely, the decoder embodies the attentive phase of FIT. It employs an attention mechanism to focus selectively on certain encoded features, integrating them to construct a complete and recognized object or scene. This selective attention in neural networks is modeled after the cognitive process, where attention binds disparate features into a singular identifiable form.

To apply this to a deep learning model, one could employ an attention mechanism that echoes the selective focus of attention. For example, convolutional layers in the attention-based encoder could detect a wide array of features within input data. At the same time, the decoder, through an attention mechanism, would prioritize and fuse these features based on their relevance to the task—such as enhancing the clarity of 3-D point clouds by filtering out noise.

By integrating this selective focus into the neural network, it could be trained to identify and highlight features that are critical for specific outcomes, thereby enhancing the model's ability to discern and prioritize information in complex datasets, much like the human visual system's proficiency in focusing on salient features in a visual scene.

III. MOTIVATION

The latest technological advancements and environmental surveys often produce unwanted noise. To address this issue, our network uses the physical properties of sensors and mapping principles to filter out noise. Our design process involved two crucial studies: 1) data analysis and 2) physical priors.

A. Data Analysis

The first study identified the essential characteristics of a dataset. The following sections will discuss the findings and key features of the study.

1) *Density:* Determining the varying density of noise points across a scene is vital to our network design because it can be challenging to distinguish between foreground and noise. The noise points are not uniformly distributed and can appear in clusters, resembling complex objects, such as trees or bushes, while others are sparsely dispersed.

2) *Localization:* The dataset also reveals another distinguishing characteristic of noise points: their localization. As mentioned, many noise points are projected at the transition point between two PIA zones. These atmospheric points exhibit a systematic global pattern that can be either dense or sparse, depending on the environmental conditions of the surveying area. Interestingly, this recurring noise pattern can be close to the objects of interest in some cases, while in others, it is found below the terrain.
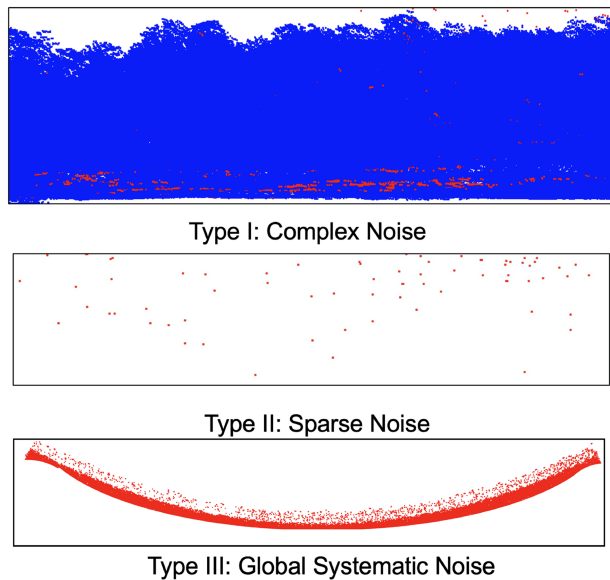
Fig. 5.   Three noise types: 1) complex; 2) sparse noise; and 3) global systematic.

3) *Distribution:* The PIA zones are narrower because of the higher PRF, which results in atmospheric points overlapping with the foreground. This phenomenon is particularly noticeable in areas with trees, bushes, buildings, power lines, and, occasionally, even the terrain. Despite this, the atmospheric points account for only a small percentage (1–3%) of the overall foreground in the dataset.

B. Physical Priors

We conducted an extensive analysis to investigate various aspects of noise data points, such as their distribution, projection near the PIA transition zone, clustering with objects of interest, and how they affect the perceived quality of the scene [58]. The range ambiguity resolution of noise points results in projecting noise on the transition among PIA zones. Hence, our network leverages this physical information, including PIA, PRF, and flight trajectory, to calculate the proximity of each point to the PIA transition zone. These proximity data are then used to generate a heatmap that indicates the probability of a point being a noise. The heatmap serves as a physical prior or external cue, enabling our network to effectively incorporate global context and filter out noise.

Based on these studies, we identified three types of noise, as illustrated in Fig. 5. This categorization motivated us to propose a design integrating physical priors to filter out noise.

1) *Complex noise:* This type of noise shares similar characteristics with trees and bushes and often appears in clusters that overlap with objects of interest.
2) *Sparse noise:* It consists of outlier and sparse noise points located far away from the terrain and objects of interest.
3) *Global systematic noise:* This noise type exhibits a dense or sparse distribution and follows a parabolic systematic

pattern localized at the transition point among PIA zones (refer to Fig. 5).

IV. METHODOLOGY

NSANet is a physical-prior-based deep neural network for noise filtering that employs a multiresolution approach and the dual-attention mechanism. Our network is designed to address all three types of noise mentioned earlier. We utilize spatial and physical-prior-based attention to reduce ambiguity in distinguishing between noise and non-noise objects, as proposed in [10]. Our design has three major components: 1) a multiresolution encoder–decoder structure, which targets sparse noise spread out as outliers; 2) a dual-attention module that focuses on physical priors or global attention to address global systematic noise; and 3) local spatial attention that targets the ambiguous boundaries between complex noise and objects of interest. In this context, we will highlight the significance of the fusion of the attention module in the encoder and decoder blocks based on the AET and FIT.

A. 3-D Noise-Filtering Neural Network

The NSANet architecture, depicted in Fig. 6, is a U-shaped multiresolution encoder–decoder structure that utilizes a 3-D voxel grid to learn deep features for noise labeling and filtering [9]. We opted for a voxel-based architecture based on a comparative study, which demonstrated that 3-D voxel-based networks outperform point-based networks (such as PointNet [7] and PointNet++ [8]) in terms of efficiency and performance for all types of noise. Point-based networks fail to cater to global systematic and complex noise and are also slow compared with voxel-based networks. One of the key reasons for selecting the voxel-based approach is its ability to deal with spatial hierarchies, which is essential for this specific problem due to the requirement of feature extraction from multiscale contexts for noise filtering.

The network encodes four feature maps of different resolutions, denoted by E_1, E_2, E_3, and E_4. The first few layers extract essential low-level features, while deeper layers extract more sophisticated features. E_n decodes feature map D_n using local spatial additive attention between D_{n-1} and E_n, followed by multiplicative attention using the physical prior probability map P_n. This process is then followed by two $3 \times 3 \times 3$ convolution operations, batch normalization, and ReLu activation. The output consists of three feature maps, i.e., D_1, D_2, D_3, each with dimensions $\frac{H}{2l} \times \frac{W}{2l} \times \frac{D}{2l} \times 32l$, where l represents the number of layers. These feature maps are trained to focus on all three types of noise at multiple receptive fields, integrate global receptive fields through global attention, and subsequently remove noise points.

1) Physical Prior Estimation (PPE) Module: The attention module in our network requires a physical prior to accurately estimating global attention. We utilize the speed of light (c) and PRF to estimate the maximum LiDAR range (R_{max}). In addition, the system takes the minimum and maximum GPS time (P_{mingt} and P_{maxgt}) recorded from the 3-D point cloud to determine the position (T_{match}) of the sensor during a specific time. The
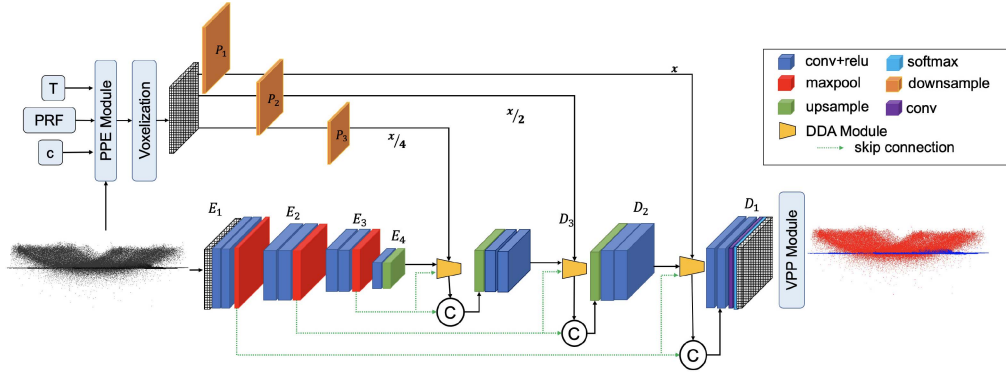
Fig. 6. NSANet-1 based on AET. It uses a PPE module to generate a PIA proximity probability map converted into a voxel grid and downsample by a factor of 2. The multiresolution encoder–decoder structure uses a DDA module to eliminate noise based on local and global contexts. The VPP module eliminates borderline misclassification of noise points.
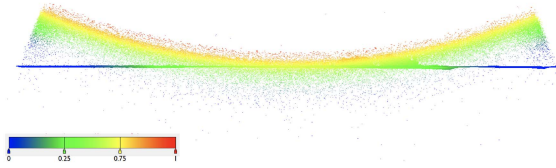

Fig. 7. Estimated physical priors: Physical prior probability is calculated using Algorithm 1. The higher the proximity to the PIA transition zone, the higher the probability of the point being part of a global systematic noise pattern, as shown with red color.

Algorithm 1: Physical Prior Estimation Module.

Input: PRF, c, P_n, T_l, N, N_T
$n = 1, 2, 3, \ldots., N, l = 1, 2, 3, \ldots., N_T \ P_{\min gt}, P_{\max gt}$

$R_{\text{MAX}} = \frac{c}{2 \times \text{PRF}}$
while $l \leq N_T$ **do**
   $T_{\text{match}} = T_l \leq P_{\min gt}$ and $T_l \geq P_{\max gt}$
end while
while $n \leq N$ **do**
   $T_m = \|P_{gt} - T_{\text{match}}\| < 1e - 2$
   $R_n = \sqrt{(P_n - T_m)^2}$
   $PIA_n = ceil(\frac{R_n}{R_{\max}})$
   $R_{\text{obs}}^{[n]} = R_n mod R_{\max}$
   $Prob_n = R_{\text{obs}}/R_{\max}$
end while
return $Prob$

---

trajectory of the flight (T_m) and 3-D point clouds (P_n) are used to estimate the range of each shot (R_n). Then, $R_{\max}$ and R_n are used to calculate their respective PIA zones (PIA$_n$). Furthermore, PIA zones, R_n, and $R_{\max}$ are utilized to compute the observed range (R_{obs}) beyond PIA zones, which is then translated into a probability map based on the distance from the transition point (Prob$_n$). The closer a point is to the transition among PIA zones, the higher the probability it is part of the parabolic systematic noise pattern, as shown in Fig. 7.

In Algorithm 1, N is the total number of raw input points and N_T is the total number of shots in flight trajectory.

2) Decoder Dual-Attention (DDA) Module: Shallow machine learning architectures or denoising autoencoders (DAEs) [59] can easily label sparse noise. However, complex noise often exhibits similar characteristics to objects of interest, making it difficult to distinguish between the two due to the distribution of points and shape variance. This ambiguity can lead to an overpopulation of false positives. We use local additive attention gates to address this issue, which handles spatial ambiguities, reduces misclassification, and suppresses irrelevant features when labeling complex noise. Global systematic noise patterns have been one of the most challenging problems in noise filtering. However, after careful study, we discovered that they only occur close to the PIA transition zone. Hence, the physical prior's heatmap from the PPE module embedded this information as global multiplicative attention. This attention mechanism focuses on global systematic noise patterns.

This novel attention model distinguishes itself from the existing attention mechanisms by seamlessly integrating both additive and multiplicative attention, thereby facilitating the imposition of local and global contexts. This model incorporates a broader context, unlike conventional approaches that solely compute similarity maps through skip connections and hidden layers. This extended context is advantageous for mitigating ambiguous boundaries in local regions and activating pertinent global regions, contributing to a more comprehensive understanding of the data

$$q_{\text{add}}^l = \sigma_1(W_x x_i^l + W_H x_i^l + b_H). \tag{4}$$

These attention gates take input from the previous layer of the decoder (D_{l-1}) for D_l, as shown in (4); $W_H x_i^l$ represents the weight of decoder tensor h_{l-1}, and b_H is bias

$$q_{\text{mul}}^l = (q_{\text{add}}^l \times W_p x_i^l). \tag{5}$$

The encoder feature map of the same level (E_2), as shown in (4), is denoted as $W_x x_i^l$ and the prior probability map of the same level (P_2) presented in (5) is denoted as $W_p x_i^l$. To calculate the local spatial attention to target complex noise, (4) is used to perform the additive operation, and then, σ_1 is
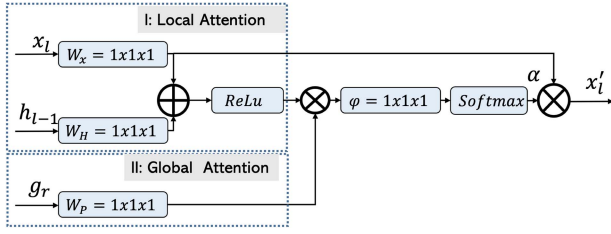
Fig. 8. Dual-attention module calculates attention coefficient based on the local and global feature maps. The local feature map from the encoder x_l and the previous layer of the decoder or encoder h_{l-1}, and global feature map comes from physical priors g_r, which are calculated based on the global positioning of each point in the point cloud from the PPE module.

Algorithm 2: Voxel Postprocessing Module.

Input: $V_n = 3D\ voxel\_grid$, $t = threshold$,
$w\_size = slidding\_window\_size$, $n = 1, 2, 3, \ldots.., N$

---

   while $n \leq N$ **do**
     If $V_n[occupancy] > 0$ and $V_n[occupancy] \geq t$ is true
     $max\_label = maxpool(V_n, w_{\text{size}})$
       If $V_n[label]$ **notequal** max_label
     $V_n[label] = max\_label$
     end If
   end If
 end while
   return $V_n^{\text{updated}} = V_n$

---

used to reflect ReLu operation $\sigma_1(x) = \max(0, x)$. The output q_{add}^l is then used to generate multiplicative attention of global physical-prior-based attention using (5)

$$q_{\text{add\_mul}}^l = \varphi^T(q_{\text{mul}}^l) + b_\varphi. \quad (6)$$

An multi-layer perception (MLP) with $1 \times 1 \times 1$ generated a dual-attention tensor with bias b_φ in (6)

$$\alpha_i^l = \sigma(q_{\text{add\_mul}}^l(x_i^l, g_r^l, h_{l-1}; \Theta_i)). \quad (7)$$

The output $q_{\text{add\_mul}}^l$ for each layer l is then normalized using sigmoid operation $\sigma_2(x) = \frac{1}{1+e^{(-x)}}$. This operation generates an attention coefficient (α), which is used to take an elementwise summation with the feature map E_2, as shown in Fig. 8 and (7).

3) Voxelization: To prepare the 3-D point cloud for input into the NSANet architecture, we represent it as a voxel grid. Raw point cloud data are preprocessed onto the voxel grid, and a mean is calculated to represent all the points within each 3-D voxel. The network also maintains a projection matrix from the voxel grid to the raw point cloud, making it easy to project labels. Choosing an appropriate voxel size provides a tradeoff between efficiency and effectiveness.

In Algorithm 2, N is the total number of voxels.

4) Voxel Postprocessing (VPP) Module: This module is an optional postprocessing component designed to refine the predictions made by NSANet. The VPP module algorithm can be seen in Algorithm 2.

1) VPP takes the input of the predicted labels of a voxel grid V_n with dimensions (Vx, Vy, Vz, C).

2) It also requires parameters such as the threshold t for the occupancy of noise points in each voxel and a sliding window size (w_x, w_y, w_z) to determine the neighborhood.

3) The module iterates over all voxels and checks if the occupancy is greater than zero but lower than the configured threshold t.

4) If the condition is met, it applies a maxpool operation with a sliding window of size $(3 \times 3 \times 3)$ on the voxel grid.

5) It checks if the maximum value of the label is equal to the label at the center of the window. For example, if the max-pooled value is 2, the module verifies if it matches the label assigned by NSANet. If a label mismatch is detected, the module assigns the max-pooled label.

6) This process is repeated for the entire voxel grid, resulting in an updated 3-D voxel grid V_n^{updated}.

V. Experiments

We conducted several experiments to validate our hypothesis. Our first ablation study compared the performance of NSANet-1, NSANet-2, and NSANet-3 with baseline while keeping two input features of occupancy points per voxel (Occ.) and mean-Z (MZ) and weighted cross-entropy (WCE) loss function. After finding optimized settings from this Ablation study, we conducted feature engineering. This suggests that employing a smaller set of features, as opposed to a limited number of manually designed geometric features, can lead to improved performance both quantitatively and in terms of efficiency. We conduct a loss- and attention-based ablation study to obtain the best NSANet configuration and optimal feature input. Furthermore, our research conducted a comparative analysis that supports our decision to utilize a voxel-based network for noise-filtering neural networks. The experiments evaluate test sets specifically designed to assess the noise class. The tables presented in this section display the best performance factors, which are highlighted in bold numbers.

A. Dataset

We employed a Teledyne Galaxy T1000 laser scanner to collect data, covering an area of 13km^2. The collected data were subsequently divided into train and test sets for our experiments. Specifically, the first 4km^2 of the dataset was allocated for testing purposes, while the remaining data were utilized to train our network.

Within the dataset, 14 distinct scenes did not overlap, encompassing over five million points on average, with a density of approximately 30 pp/m^2. To establish the ground truth, Teledyne Optech employed Terrasolid point cloud processing software [60] to label the data. These labels were generated based on extensive domain knowledge and technical expertise.

Our training dataset had two classes, i.e., noise and non-noise, which allowed us to train our network effectively.

B. Experimental Configuration

The dataset undergoes a split of 70% and 30% for training and testing, respectively, with the training set further divided
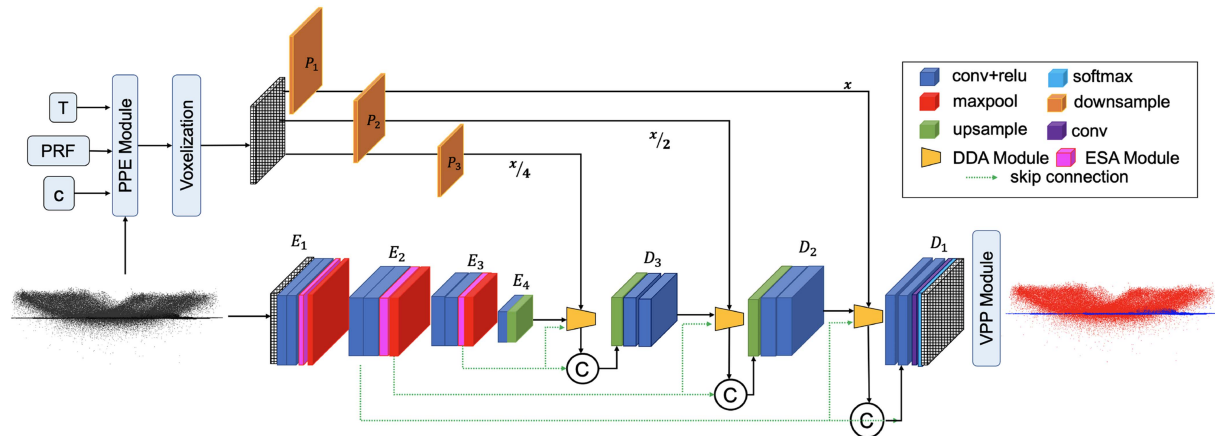
Fig. 9.   NSANet-2 adapted FIT by employing the attention module in the encoder as well. The first variation contains only local attention in encoder layers called the encoder single attention (ESA) module.

into training and validation sets (80% and 20%, respectively). Subscenes based on the GPS time of the flight line are generated within each scene, and a voxel grid of 64 m × 64 m × 64 m is created for each subscene. Features include mean absolute elevation (MZ) and the number of occupied points (Occ.). NSANet outputs confidence scores for noise and non-noise classes, with the final prediction determined by the highest confidence score. Postprocessing techniques enhance accuracy.

NSANet training involves 90 epochs on two GPU RTX 6000 units, utilizing the Adam optimizer with a learning rate of $\alpha = 0.0001$ and a 0.8 decay rate for every 100 000 steps through a staircase learning rate schedule. Inference takes 2–3 min, and precision, recall, and F1-score assess performance. PointNet and PointNet++ use a 1-m³ partition size, training for 90 epochs on two GPU RTX 6000 units with a learning rate of $\alpha = 0.001$, and a 0.7 decay rate for every 200 000 steps via a staircase learning rate schedule. SVM training involves voxelization, grouping 100 points using voxels, and extracting features like the number of neighboring points, mean elevation, and geometric features (omnivariance, eigentropy, anisotropy, planarity, linearity, surface variation, sphericity, and verticality). The SVM is trained with a soft-margin radial basis function kernel of gamma = 0.05.

C. Evaluation Matrices

We selected precision, recall, and F1-scores to evaluate our network's performance as noise is in the minority class. The precision determines the total number of predicted noise or non-noise out of the total instances of the class, as shown in (9). Recall measures correctly predicted samples out of the total number of predicted samples of the specific class, as shown in (8). The F1-score shows the balance between precision and recall, as shown in (10). The reason for selecting these evaluation metrics is that the noise-based evaluation metrics, such as peak signal-to-noise ratio or structural similarity index, do not apply to point clouds as these are for image denoising. However, distance-based evaluation metrics, such as Chamfer distance, earth mover's distance, and fidelity metrics, are insufficient for

evaluating the performance of binary classification of noise and non-noise points

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (8)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (9)$$

where TP is true positive, FN is false negative, and FP is false positive

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \qquad (10)$$

VI. RESULTS AND DISCUSSION

A. NSANet Ablation Study

Our ablation study introduced a dual-attention module in the encoder and the decoder based on our interpretation of FIT and AET. Through a series of experiments, we validated that the implementation of NSANet-1, as depicted in Fig. 6, which focuses on utilizing attention during the decoding stage, outperformed the variations of NSANet-2 and NSANet-3 illustrated in Figs. 9 and 10. This outcome confirms that the network effectively leverages the principles of attentional theory as exemplified by AET, thus experimentally validating its usability. The results presented in Table I highlight that incorporating attention mechanisms during the feature interpretation stage to filter noise is a superior strategy for addressing this problem. It also validates the human vision principle of paying attention to the decoding stage instead of the data extraction stage, as it gives balanced results. At the same time, NSANet-2 has more omission errors, and NSANet-3 has more commission errors. Noise filtering does not benefit from overly aggressive or conservative networks.

B. Feature Engineering

The input features for deep learning networks are crucial for learning. Initially, the experiment is configured to select a palette of features, including Occ. and MZ, for the NSANet-1 ablation study. The mean elevation for each voxel was calculated based
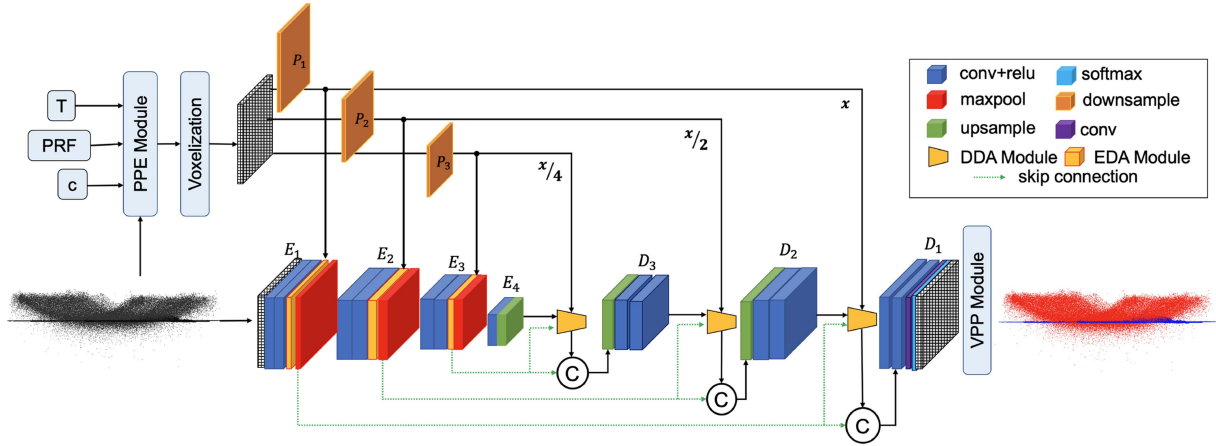
Fig. 10.   NSANet-3 adapted FIT by employing the attention module in the encoder as well. This variation contains dual attention in encoder layers called the encoder dual-attention (EDA) module.
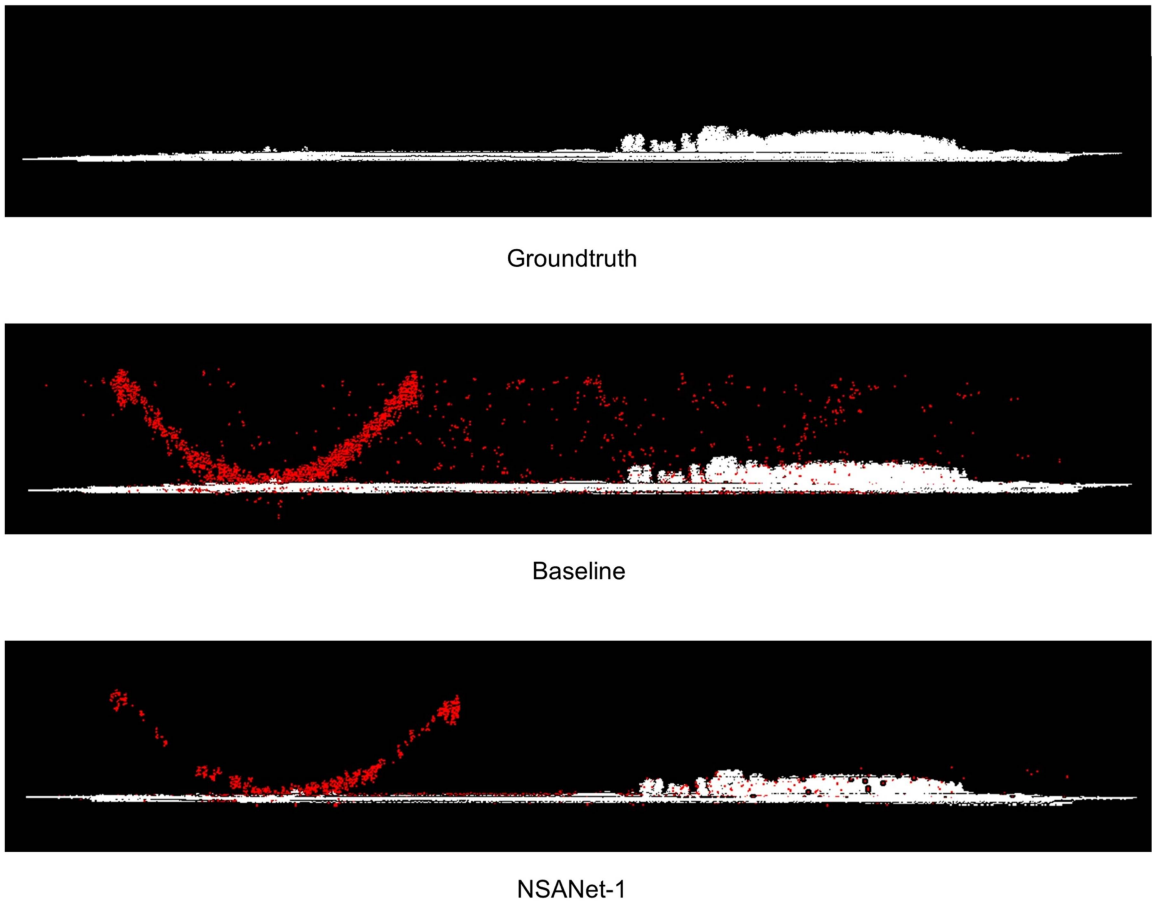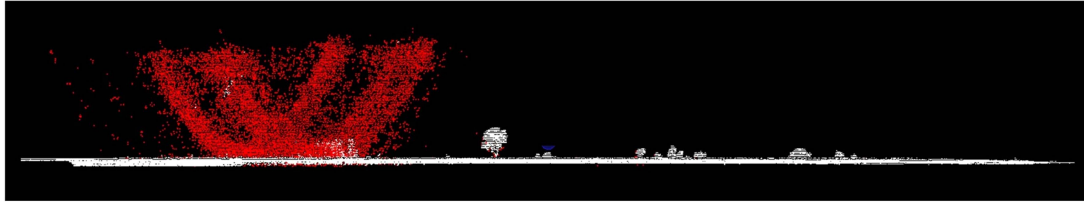

Fig. 11.   Visualization shows scenes with high vegetation comparing baseline performance with WCE loss and NSANet-1-variation-employed AET and WCE loss. Red shows the misclassification of noise points, and white reflects non-noise points.

on the z values of the points within the voxel, enhancing the network's ability to learn the vertical geometry of the scene (see Table II). For the proposed NSANet-1, which incorporates physical priors and spatial attention, training was conducted using a single channel, surpassing the performance of the baseline, achieving a recall of 72.5%, a precision of 85.4%, and an F1-score of 0.789%. Subsequently, MZ was added to the feature channels alongside (Occ.), which improves the performance on recall by 14.8%, precision by 7.5%, and F1-score by 0.111%. The final results in Table II highlight the importance of absolute elevation as a discriminating feature, especially due to the distinctive noise distribution in the elevation dimension. Most of
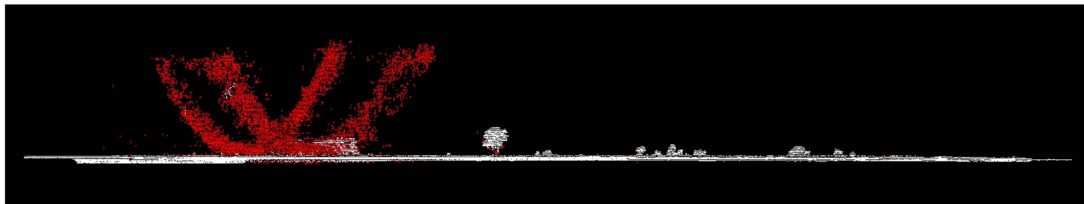
Fig. 12. Visualization shows scenes with low vegetation comparing baseline performance with WCE loss and NSANet-1-variation-employed AET and WCE loss. Red shows the misclassification of noise points, and white reflects non-noise points.

TABLE I
ABLATION STUDY OF MODULES FOR NSANET TO COMPARE THE PERFORMANCE OVER RECALL, PRECISION, AND F1-SCORE

Methods	Fusion	Attention	Features		Loss	Recall	Precision	F1-Score
			Occ.	MZ				
Baseline	✗	✗	✓	✓	WCE	77.92%	91.43%	0.841
NSANet-1	AET	DDA	✓	✓	WCE	87.30%	92.90%	**0.900**
NSANet-2	FIT	ESA+DDA	✓	✓	WCE	83.31%	**93.60%**	0.882
NSANet-3	FIT	EDA+DDA	✓	✓	WCE	**92.12%**	76.69%	0.856

TABLE II
ABLATION STUDY ON IMPORTANCE OF FEATURE ENGINEERING FOR OPTIMIZED NSANET-1 OVER RECALL, PRECISION, AND F1-SCORE

Methods	Features					Fusion	Loss	Recall	Precision	F1-Score
	Occ.	MZ	INS	NR	R		WCE			
NSANet-1	✓	✗	✗	✗	✗	AET	✓	72.52%	85.41%	0.789
NSANet-1	✓	✓	✗	✗	✗	AET	✓	**87.30%**	**92.91%**	**0.900**
NSANet-1	✓	✓	✓	✓	✓	AET	✓	**87.30%**	83.52%	0.855

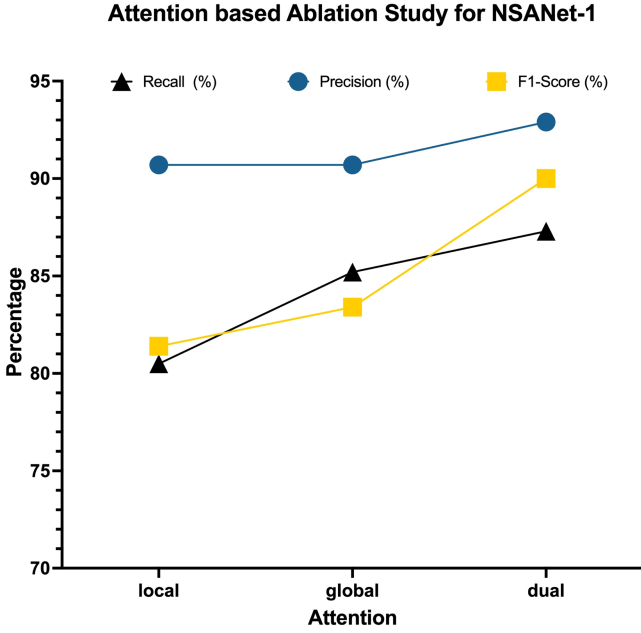Occ: Number of occupancy points, MZ: mean elevation, INS: intensity, R: range, and NR: number of returns.

Fig. 13. Comparison of attention modules and their impact on F1-score, recall, and precision for noise filtering.
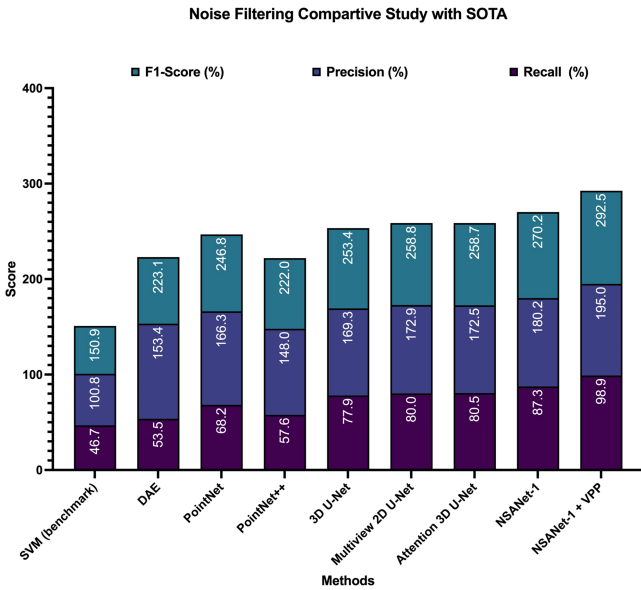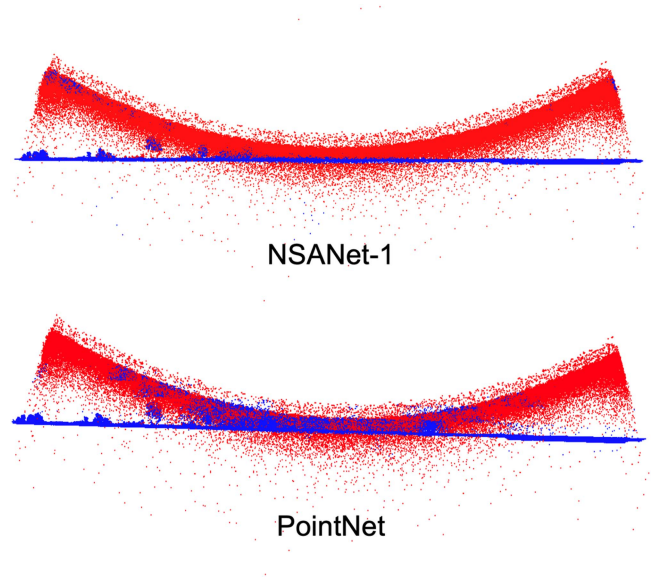

Fig. 15. Comparison of point- and voxel-based networks. Red represents noise points, and blue represents regular points. PointNet has more misclassification of noise points closer to the ground than regular points.

Fig. 14. Cumulative score of SOTA for noise filtering. It shows that the NSANet outperforms existing SOTA.
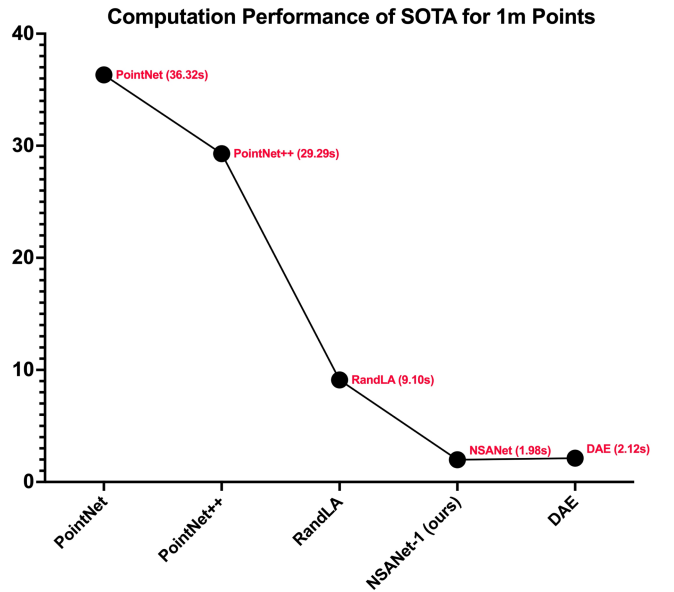
Fig. 16. Comparison of SOTA based on inference time for 1 million points.

the noise annotation is performed through 2-D vertical profiling. Thus, the feature combination of mean elevation and occupancy points will be used in all the NSANet-1 experiments.

C. Loss Functions

The research employed the cross-entropy (CE) loss, which is commonly used for segmentation problems

$$\text{CE} = \left[-\sum_{i=1}^{C} y_i \log(p_i) \right]. \tag{11}$$

In (11), y represents the true label, p is the predicted probability for the positive class in binary classification, p_i is the predicted probability for class i in multiclass classification, and C is the number of classes. Then

$$\text{WCE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} w_c \cdot y_{i,c} \cdot \log(p_{i,c}) \tag{12}$$

where N represents the total number of samples, C denotes the number of classes, w_c signifies the weight assigned to class c, $y_{i,c}$ is the true label for sample i and class c, and $p_{i,c}$ is the predicted probability for sample i and class c. However,

TABLE III
ABLATION STUDY ON IMPORTANCE OF LOSS FUNCTION OVER RECALL, PRECISION, AND F1-SCORE

Methods	Loss			Fusion	Features		Recall	Precision	F1-Score
	CE	FL	WCE		Occ.	MZ			
NSANet-1	✓	✗	✗	AET	✓	✓	72.11%	90.70%	0.814
NSANet-1	✗	✓	✗	AET	✓	✓	76.21%	90.70%	0.834
NSANet-1	✗	✗	✓	AET	✓	✓	**87.32%**	**92.91%**	**0.900**

CE: cross-entropy loss, FL: focal loss, and WCE: weighted cross entropy.

TABLE IV
ABLATION STUDY OF GLOBAL, LOCAL, AND DUAL-ATTENTION MODULES IN THE REFERENCE TO THE DDA MODULE IN FIG. 8 TO COMPARE PERFORMANCE ADVANTAGE OVER RECALL, PRECISION, AND F1-SCORE

Methods	DDA Module		Fusion	Features		Loss	Recall	Precision	F1-Score
	local	global		Occ.	MZ	WCE			
NSANet-1	✓	✗	AET	✓	✓	✓	80.51%	92.00%	0.862
NSANet-1	✗	✓	AET	✓	✓	✓	85.22%	91.60%	0.884
NSANet-1	✓	✓	AET	✓	✓	✓	**87.30%**	**92.90%**	**0.900**

TABLE V
COMPARISON OF NSANET FOR NOISE FILTERING WITH SVM, DAE, 3-D U-NET, POINTNET++, AND POINTNET OVER RECALL, PRECISION, AND F1-SCORE FOR NOISE CLASS

Methods	input	Recall	Precision	F1-Score
SVM (benchmark)	segment	46.78%	54.05%	50.1%
DAE [59]	voxel	53.57%	**99.90%**	69.7%
PointNet [7]	point	68.29%	98.19%	80.5%
PointNet++ [8]	point	57.65%	90.42%	74.0%
3-D U-Net [9]	voxel	77.92%	91.43%	84.1%
Att-U-Net [10]	voxel	80.51%	92.00%	86.2%
Multiview 2D-U-Net	views	80.00%	92.90%	85.9%
NSANet-1	voxel	**87.30%**	92.90%	**90.0%**

TABLE VI
COMPARISON OF NSANET FOR COMPUTATIONAL PERFORMANCE WITH SOTA

Methods	Training time per epoch	Parameters	Inference time
DAE	**850s**	**0.2m**	2.12s
PointNet	1600s	0.8m	38.32s
PointNet++	1457s	0.97m	29.29s
RandLA	1200s	1.24m	9.10s
NSANet-1	1000s	4.0m	**1.98s**

the specific distribution and characteristics of the noise and non-noise classes in data have not been effectively addressed by FL exhibiting recall decrease by 11.1%, precision by 2.2%, and F1-score by 6.6%. Therefore, WCE loss significantly outperformed FL, confirming its superiority in this problem domain.

D. Attention Ablation Study

Furthermore, another ablation study assessed the significance of global, local, and dual-attention modules. The findings presented in Table IV and Fig. 13 demonstrate that integrating physical priors yields a relatively higher performance improvement of 85.2% recall, 91.6% precision, and 0.884% F1-score compared to using only local attention, which shows a performance decrease of 4.71% recall, 0.4% precision, and 0.022 F1-score. However, the dual-attention module plays a critical role in effectively eliminating noise from the point cloud, facilitating comprehensive scene understanding with an F1-score of 0.900. These experiments have effectively validated the importance of physics-based deep learning approaches for computer vision tasks. This ablation study shows that despite local attention's ability to deal with ambiguous territories between noise and

this specific problem necessitated using a WCE loss, as shown in (12). This was due to the significant class imbalance in the dataset, making it crucial to assign appropriate weights. The introduction of class weights proved highly beneficial and led to a substantial improvement in performance, as demonstrated in Table III. The results show that the performance for the noise class was best when WCE is used with the recall of 87.3%, the precision of 92.9%, and the F1-score of 90.0%. To explore alternative optimal loss functions, a few tests were conducted using focal loss (FL), as shown in the following equation, to assess its impact on performance:

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (13)$$

where p_t is the predicted probability of the true class, α_t is the FL balancing factor, and γ is the focusing parameter. In addition,

TABLE VII
ABLATION STUDY BASED ON THE VPP MODULE AND COMPARED PERFORMANCE OVER RECALL, PRECISION, AND F1-SCORE

| Methods | VPP | Fusion | Features | | Recall | Precision | F1-Score |
			Occ.	MZ			
NSANet-1	✗	AET	✓	✓	87.30%	92.90%	90.0%
NSANet-1	✓	AET	✓	✓	**98.90%**	**96.10%**	**97.5%**

objects of interest, it fails to deal with global systematic patterns. Therefore, global attention fused with local attention works best for removing noise. Qualitative analysis of attention fusion shows that every attention module plays an important role in removing noise points. Point-based graph shows how global attention improves results more than local attention, but an overall performance boost is achieved by including both modules as dual attention.

E. Comparative Study

We conducted a comparative study based on an extensive literature review. The selection of existing deep learning networks for noise filtering was not chosen due to their inability to deal with 3-D large-scale point clouds. We selected the models used for semantic segmentation of the 3-D point cloud due to their ability to deal with large-scale data and multiscale objects and integrate larger contexts. Considering these limitations, various models, including SVMs, DAE, PointNet, PointNet++, 3-D UNet, Attention 3-D UNet, and multiview-based 2-D UNet, as shown in Table V, were selected for noise filtering. The results of each model are summarized as follows.

1) The SVM was trained to balance preserving objects of interest and eliminating noise points. It achieved an F1-score of 0.501 for the noise class.
2) DAE, which focuses on denoising reconstruction, outperformed SVM with a relatively higher F1-score of 0.697 for the noise class
3) PointNet [7], a pioneering network that directly processes 3-D point cloud data, performed better than DAE but struggled with complex noise and lacked global context, multiresolution, and local contextual features and shows the F1-score of 0.805. Fig. 15 shows the superior performance of the voxel-based network.
4) PointNet++ [8] was also tested but did not yield satisfactory results (F1-score 0.740).
5) The 3-D U-Net model with WCE loss was selected as the baseline. It utilized a multiresolution encoder–decoder structure to incorporate more context and exhibited exponentially improved performance on the noise class by exhibiting a recall of 77.9% and F1-score of 0.841.
6) It further compared the performance of attention 3-D U-Net and multiview-based 2-D U-Net. The 2-D U-Net, which took two views as input, showed comparable performance to the attention 3-D U-Net by exhibiting the

F1-score of 0.859 and 0.862. The 2-D U-Net has a limitation as the view selection can be tedious and dependent on the dataset. However, attention 3-D U-Net exhibited aggressive noise classification, misclassifying non-noise as noise, while the 3-D U-Net with CE loss demonstrated generally poor performance.

Based on the insights gained from these experiments, we concluded that our system, which incorporates global and local contexts through attention mechanisms embedded in the network, works effectively. This approach led to significant performance enhancements, as demonstrated in Table IV. Unlike other networks that displayed better precision but lower recall and F1-scores, our model achieved a balanced performance across all metrics, as shown in Fig. 14.

F. Computational Performance

Our comparative evaluation underscores the superior computational efficiency of the NSANet architecture, a voxel-based neural network. This assessment, illustrated in Fig. 16, benchmarks the performance of NSANet against other state-of-the-art (SOTA) networks by measuring the time it takes to infer information from 1 million data points. Table VI showcases our findings, with NSANet significantly outpacing its counterparts at an inference speed of 1.98 s for the given data size. Point-based networks, such as PointNet and PointNet++, lag behind, with inference times of 36.32 and 29.29 s, respectively, while RandLA completes the task in 9.10 s. NSANet's rapid inference capabilities could be beneficial for a range of practical applications that require efficient noise filtering.

In this study, the architecture of NSANet has been refined from its original baseline, resulting in a reduction of parameters when compared to the standard 3-D U-Net. However, it employs a relatively high number of trainable parameters at 4.0 million. This figure is indicative of the network's complexity and the computational resources required, which may present a limitation in certain use cases. Despite this, the architecture offers a significant advantage in processing speed and noise-filtering capability, which are often critical considerations in practical applications. The optimization of the NSANet design balances this tradeoff, leveraging a higher parameter count to achieve a level of accuracy and efficiency that may justify the increased computational demand in scenarios where rapid processing is prioritized.

Nevertheless, future work could focus on parameter reduction strategies without compromising the network's performance to enhance its applicability in resource-constrained environments.

It is worth mentioning that the inclusion of RandLA in our comparison provides a broader context for evaluating NSANet's performance. RandLA's noise-filtering capabilities were not explicitly tested in our study due to the limitations of point-based networks in noise management and their challenges to classify overlapping classes in semantic segmentation tasks. However, it remains pertinent to demonstrate NSANet's competitive advantage in terms of inference efficiency.

G. G. Error Analysis and VPP Module

Our research has shown advancements in noise filtering, surpassing the SOTA methods, as illustrated in Figs. 11 and 12. It is important to highlight that our work draws inspiration from semantic segmentation and denoising research. Filtering noise from large-scale 3-D point cloud data captured by airborne LiDAR remains a challenging problem with limited existing solutions. However, there are certain scenarios where our network encounters difficulties in filtering noise, particularly in cases where the distribution of objects of interest closely resembles noise or overlaps with it. These challenging instances serve as hard examples and contribute to a performance gap. To address this, we have incorporated a VPP module, which significantly enhances the network's performance by 11.6% (recall), 3.2% (precision) and 7.5% (F1) by considering neighboring voxel's confidence. Detailed information regarding the impact of VPP can be found in Table VII.

VII. CONCLUSION

This study introduced an innovative dual-attention NSANet model that leverages physical priors and local spatial attention for effective noise filtering. Through our extensive experimental ablation study, we demonstrated that incorporating global and local contexts as attention mechanisms can effectively address the challenges posed by overlapping objects and remove systematic global patterns and sparse noise. Furthermore, our study provided evidence that integrating global context enhances scene understanding. These experiments validated our conceptual inspiration and confirmed the hypothesis behind our two-stage-based FIT, which incorporates physical priors into the network. Importantly, our results showcased that this integration significantly improves noise-filtering performance. NSANet exhibited impressive capabilities and efficiency in handling large-scale point clouds captured by airborne LiDAR scanners. Unlike other models, our approach consistently performed well across all evaluation metrics. Moving forward, our future research will focus on exploring point-based neural networks and advancing the learning of physical priors to enhance noise removal capabilities further.

ACKNOWLEDGEMENT

REFERENCES

[1] N. El-Sheimy, "An overview of mobile mapping systems," in *Proc. Int. Conf. Glob. Spatial Data Infrastruct.*, 2005, pp. 1–14.

[2] G. Petrie, "Mobile mapping systems: An introduction to the technology," *Geoinformatics*, vol. 13, pp. 32–43, Jan. 2010.

[3] M. Lemmens, *Geo-Information: Technologies, Applications and the Environment*. Berlin, Germany: Springer, 2011.

[4] D. Hartsell, P. LaRocque, and J. Tripp, "Galaxy: A new state of the art airborne lidar system," *Lidar News*, vol. 1, no. 1, pp. 26–31, 2010.

[5] M. J. Rakotosaona, V. La Barbera, P. Guerrero, N. J. Mitra, and M. Ovsjanikov, "PointCleanNet: Learning to denoise and remove outliers from dense point clouds," *Comput. Graph. Forum*, vol. 39, no. 1, pp. 185–203, 2020.

[6] B. Goyal, A. Dogra, S. Agrawal, B. S. Sohi, and A. Sharma, "Image denoising review: From classical to state-of-the-art approaches," *Inf. Fusion*, vol. 55, pp. 220–244, 2020.

[7] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 77–85, doi: 10.1109/CVPR.2017.16.

[8] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet: Deep hierarchical feature learning on point sets in a metric space," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 5105–5114.

[9] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2016, pp. 424–432.

[10] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[11] S. Luo and W. Hu, "Score-based point cloud denoising: Learning gradient fields for point cloud denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 4563–4572, doi: 10.1109/ICCV48922.2021.00454.

[12] S. Luo and W. Hu, "Differentiable manifold reconstruction for point cloud denoising," in *Proc. 28th ACM Int. Conf. Multimedia*, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1330–1338, doi: 10.1145/3394171.3413727.

[13] M. Gadelha, R. Wang, and S. Maji, "Deep manifold prior," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Montreal, BC, Canada, 2021, pp. 1107–1116, doi: 10.1109/ICCVW54120.2021.00130.

[14] C. Huang, R. Li, X. Li, and C.-W. Fu, "Non-local part-aware point cloud denoising," 2020, *arXiv:2003.06631*.

[15] W. Hu, X. Gao, G. Cheung, and Z. Guo, "Feature graph learning for 3D point cloud denoising," *IEEE Trans. Signal Process.*, vol. 68, pp. 2841–2856, 2020, doi: 10.1109/TSP.2020.2978617.

[16] F. Pistilli, G. Fracastoro, D. Valsesia, and E. Magli, "Learning graph-convolutional representations for point cloud denoising," 2020, *arXiv:2007.02578*.

[17] H. P. Casajus, T. Ritschel, and T. Ropinski, "Total denoising: Unsupervised learning of 3D point cloud cleaning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Korea (South), 2019, pp. 52–60, doi: 10.1109/ICCV.2019.00014.

[18] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[19] J. Duncan and G. Humphreys, "Beyond the search surface: Visual search and attentional engagement," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 18, no. 2, pp. 578–88, May 1992, doi: 10.1037//0096-1523.18.2.578.

[20] K. C. Slatton, W. E. Carter, R. L. Shrestha, and W. Dietrich, "Airborne laser swath mapping: Achieving the resolution and accuracy required for geosurficial research," *EOS Trans.*, vol. 88, no. 17, pp. 193–200, 2007.

[21] P. Rieger and A. Ullrich, "Resolving range ambiguities in high-repetition rate airborne lidar applications," *Proc. SPIE*, vol. 8186, 2011, Art. no. 81860A.

[22] R. B. Roth and J. L. Thompson, "Practical application of multiple pulse in air (MPIA) LiDAR in large-area surveys," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. XXXVII, no. B1, pp. 183–188, 2008.

[23] Y. Ba, G. Zhao, and A. Kadambi, "Blending diverse physical priors with neural networks," 2019, *arXiv:1910.00201*.

[24] R. Thoreau, L. Risser, V. Achard, B. Berthelot, and X. Briottet, "p³ VAE: A physics-integrated generative model. application to the semantic segmentation of optical remote sensing images," 2023, *arXiv:2210.10418*.

[25] G. Ju, Y. Choi, D. Lee, J. H. Paik, G. Hwang, and S. Lee, "Self-supervised dehazing network using physical priors," in *Proc. 16th Asian Conf. Comput. Vis.*, 2023, pp. 290–305.

[26] J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar, "Integrating scientific knowledge with machine learning for engineering and environmental systems," *ACM Comput. Surv.*, vol. 55, no. 4, Nov. 2022, Art. no. 66, doi: 10.1145/3514228.

[27] O. Schall, A. Belyaev, and H.-P. Seidel, "Robust filtering of noisy scattered point data," in *Proc. Eurographics/IEEE VGTC Symp. Point-Based Graph.*, 2005, pp. 71–144.

[28] S. Ronnback and A. Wernersson, "On filtering of laser range data in snowfall," in *Proc. IEEE 4th Int. Conf. Intell. Syst.*, 2008, pp. 17-33–17-39.

[29] N. Charron, S. Phillips, and S. L. Waslander, "De-noising of LiDAR point clouds corrupted by snowfall," in *Proc. 15th Conf. Comput. Robot Vis.*, 2018, pp. 254–261.

[30] P. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley Interdiscipl. Rev. Data Mining Knowl. Discov.*, vol. 1, pp. 73–79, 2011.

[31] B. Zou, H. Qiu, and Y. Lu, "Point cloud reduction and denoising based on optimized downsampling and bilateral filtering," *IEEE Access*, vol. 8, pp. 136316–136326, 2020.

[32] C. Lv and M. Li, "Point cloud denoising algorithm based on noise classification," in *Proc. Int. Conf. Culture-Oriented Sci. Technol.*, 2020, pp. 123–127.

[33] E. Mattei and A. Castrodad, "Point cloud denoising via moving RPCA: MRPCA," *Comput. Graph. Forum*, vol. 36, pp. 123–137, Nov. 2016.

[34] F. Zaman, Y. Wong, and B. Ng, "Density-based denoising of point cloud," 2016, *arXiv:1602.05312*.

[35] Q. Kang, G. Huang, and S. Yang, "A gross error elimination method for point cloud data based on Kd-tree," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. XLII-3, pp. 719–722, Apr. 2018.

[36] Y. Schoenenberger, J. Paratte, and P. Vandergheynst, "Graph-based denoising for time-varying point clouds," in *Proc. 3DTV-Conf. True Vis.—Capture Transmiss. Display 3-D Video*, 2015, pp. 1–4.

[37] C. Dinesh, G. Cheung, I. V. Bajic, and C. Yang, "Fast 3D point cloud denoising via bipartite graph approximation and total variation," 2018, *arXiv:1804.10831*.

[38] C. Dinesh, G. Cheung, and I. V. Bajic, "3D point cloud denoising via bipartite graph approximation and reweighted graph Laplacian," 2018, *arXiv:1812.07711*.

[39] J. Digne and C. de Franchis, "The bilateral filter for point clouds," *Image Process. On Line*, vol. 7, pp. 278–287, Oct. 2017.

[40] S. Y. Hua, Z. Xu-qing, N. Xue-feng, Y. guo-dong, and Z. Ji-Kai, "Denoising algorithm of airborne LIDAR point cloud based on 3D grid," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 10, pp. 85–92, Mar. 2017.

[41] W. Hu, Q. Hu, Z. Wang, and X. Gao, "Dynamic point cloud denoising via manifold-to-manifold distance," *IEEE Trans. Image Process.*, vol. 30, pp. 6168–6183, 2021, doi: 10.1109/TIP.2021.3092826.

[42] R. Gao, J. Park, X. Hu, S. Yang, and K. Cho, "Reflective noise filtering of large-scale point cloud using multi-position LiDAR sensing data," *Remote Sens.*, vol. 13, no. 16, 2021, Art. no. 3058, doi: 10.3390/rs13163058.

[43] M. A. Irfan and E. Magli, "3D point cloud denoising using a joint geometry and color k-NN graph," in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 585–589.

[44] Y. Wang and H.-Y. Feng, "Outlier detection for scanned point clouds using majority voting," *Comput.-Aided Des.*, vol. 62, pp. 31–43, 2015.

[45] Y. Duan, C. Yang, H. Chen, W. Yan, and H. Li, "Low-complexity point cloud denoising for LiDAR by PCA-based dimension reduction," *Opt. Commun.*, vol. 482, 2021, Art. no. 126567.

[46] M. Zeybek, "Inlier point preservation in outlier points removed from the ALS point cloud," *J. Indian Soc. Remote Sens.*, vol. 49, pp. 2347–2363, Jun. 2021.

[47] Y. Regaya, F. Fadli, and A. Amira, "Point-denoise: Unsupervised outlier detection for 3D point clouds enhancement," *Multimedia Tools Appl.*, vol. 80, no. 18, pp. 28161–28177, Jul. 2021, doi: 10.1007/s11042-021-10924-x.

[48] D. Zhang, X. Lu, H. Qin, and Y. He, "Pointfilter: Point cloud filtering via encoder-decoder modeling," 2020, *arXiv:2002.05968*.

[49] C. Duan, S. Chen, and J. Kovacevic, "3D point cloud denoising via deep neural network based local surface estimation," 2019, *arXiv:1904.04427*.

[50] V. Sarode, A. Dhagat, R. A. Srivatsan, N. Zevallos, S. Lucey, and H. Choset, "MaskNet: A fully-convolutional network to estimate inlier points," 2020, *arXiv:2010.09185*.

[51] R. Sharma, T. Schwandt, C. Kunert, S. Urban, and W. Broll, "Point cloud upsampling and normal estimation using deep learning for robust surface reconstruction," 2021, *arXiv:2102.13391*.

[52] R. Gao, M. Li, S.-J. Yang, and K. Cho, "Reflective noise filtering of large-scale point cloud using transformer," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 577, doi: 10.3390/rs14030577.

[53] X. Wang, X. Fan, and D. Zhao, "PointFilterNet: A filtering network for point cloud denoising," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1276–1290, Mar. 2023.

[54] H. Gong et al., "Multiscale information fusion for hyperspectral image classification based on hybrid 2D-3D CNN," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2268, doi: 10.3390/rs13122268.

[55] M. Guo et al., "Attention mechanisms in computer vision: A survey," 2021, *arXiv:2111.07624*.

[56] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.

[57] Y. Yeshurun, "Attentional mechanisms in computer vision," *Signal Process. Appl. Artif. Vis.*, pp. 43–52, 1997, doi: 10.1016/B978-012444816-2/50006-X.

[58] M. Jameela, L. Chen, A. Sit, J. Yoo, C. Verheggen, and G. Sohn, "Simulation-based data augmentation using physical priors for noise filtering deep neural network," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. XLIII-B2-2020, pp. 247–254, Aug. 2020.

[59] A. Palla, D. Moloney, and L. Fanucci, "Fully convolutional denoising autoencoder for 3D scene reconstruction from a single depth image," in *Proc. Int. Conf. Syst. Informat.*, 2017, pp. 566–575.

[60] "Terrasolid", Accessed: Nov. 11, 2022. [Online]. Available: https://terrasolid.com

Maryam Jameela received the bachelor's and master's degrees in information technology from the National University of Science and Technology, Islamabad, Pakistan, in 2013 and 2017, respectively. She is currently working toward the Ph.D. degree with York University, Toronto, ON, Canada, specializing in remote sensing, computer vision, and machine learning.

She is also a Research Scientist with Teledyne Optech, Vaughan, ON, where she focuses on airborne remote sensing 3-D point cloud analysis and deep networks. With nine years of experience in programming, web/software design, and research, she has worked in various sectors, including public, private, and semipublic organizations. She actively collaborates with researchers from different disciplines and has contributed to multidisciplinary projects in remote sensing, computer vision, and large-scale point cloud analysis. She has made significant contributions to renowned research initiatives such as Natural Sciences and Engineering Research Council of Canada (NSERC)'s Collaborative Research and Training Experience program and Collaborative Research and Development Grant—3-D Mobility Mapping Artificial Intelligence in collaboration with NSERC and Teledyne Optech.

Gunho Sohn (Member, IEEE) received the B.Sc. and M.Sc. degrees in astronomy and space science from Yonsei University, Seoul, South Korea, in 1993, and the Ph.D. degree in geomatics engineering from University College London, London, U.K., in 2004.

He is currently an Associate Professor with the Department of Earth and Science and Engineering and the Inaugural Director of the Mobility Innovation Centre, Lassonde School of Engineering, York University, Toronto, ON, Canada. Before joining York University in 2008, he was a Research Scientist with Satellite R&D Division, Samsung Aerospace Industry, Ltd., Seoul, from 1993 to 1999. He is actively leading a multidisciplinary research team in spatial data analytics, computer vision, and machine learning. The research group has extensive experience in reconstructing a large-scale built environment using autonomous vehicles and mapping, with applications to various areas of urban digital twin technologies. His research group has participated in many large-scale research projects (Canada Foundation for Innovation, Natural Sciences and Engineering Research Council of Canada (NSERC)'s Collaborative Research and Training Experience program, Ontario Research Fund—Research Excellence, and Ontario Research Fund—Infrastructure), through which he has collaborated with multiple universities, international institutions, and industrial partners. His pioneering works of digital twining urban systems have been successfully commercialized through his industrial collaborators. He has authored or coauthored more than 150 peer-reviewed publications.

Dr. Sohn was the recipient of the Best Thesis Award from the British Remote Sensing and Photogrammetry Society in 2004. He was a two-time recipient of York's Research Leadership Award and the PEO York Chapter Engineering Research Project of the Year Award.