Machine Learning-Enhanced Standard Deviation to Detect and Handle of Outlier in the 3D Point-Cloud Data

1st Mohammad Rasoul Tanhatalab
Dep. of Information Engineering and Computer Science
Trento University, Italy
mohammad.tanhatalab@unitn.it

2nd Manuel Forrer
ferrisol Co
Graz, Austria
m.forrer@ferrisol.com

3rd Michael Nelz
Nelo Intelligence
Eichenau, Germany
michael.nelz@nelo-intelligence.com

***Abstract*—Outlier removal is an important step in 3D point cloud data processing. Various methodologies for outlier removal will be analyzed shortly in this paper by describing a range of statistical and machine-learning techniques. Thus, the intent is to analyze different techniques and evaluate their performance in order to justify unique advantages and limitations of each technique.**

These limitations and advantages advance the discussion by introducing a new approach combining statistical and machine learning techniques on a real-world dataset. It not only identifies the spatial position of outliers but also enhances the measurement precision at those specific points. This work has a lot of significance from many aspects for researchers, engineers, and practitioners. Especially in robotics, industrial processes, augmented reality, and autonomous driving, the nature of 3D point cloud data requires sophisticated outlier handling. We believe our work will serve as a reference to optimize data quality in these dynamic applications under development.

***Index Terms*—3D point cloud, machine learning, outlier detection, statistic, signal processing**

I. Introduction

In recent times, 3D point cloud data has gained much importance in several fields like robotics, augmented reality, industrial processes, environmental monitoring, and autonomous vehicles. One of the major challenges in the processing of 3D point cloud data in these areas deals with outliers.Outliers refer to points that deviate significantly from the expected distribution of the data and can have a negative impact on downstream algorithms. They come from improper measurement.

Different traditional and modern methods have been proposed for 3D point cloud outlier removal. Traditional statistical-based methods, such as the mean-shift algorithm, principal component analysis (PCA), kernel density estimation (KDE), local outlier factor (LOF), standard deviation (SD), elliptical envelope, and feature bagging; and machine learning such as isolation forest, cluster DBSCAN, K-nearest neighbors (KNN), convolutional neural networks (CNNs) methods, and deep autoencoders techniques are commonly being used. Most of these techniques are afflicted by their underlying normal distribution assumption of data, which might be unrepresentative of real-world environments. The most applicable outlier removal techniques of 3D point clouds are reviewed in this paper, categorized into statistical-based and machine learning-based methods. Each presented method has its advantages and limitations, while it uses the respective application of the method. A new machine learning method with high accuracy is provided for outlier detection. Finally, we conclude the review with the future directions of the research on the removal of outliers in 3D point clouds.

II. Literature Review

It proposes a novel approach to outlier detection using a hierarchical spatial verification scheme. The proposed method shows outstanding performance in place recognition and loop closure detection compared with the state-of-the-art solutions on five popular public datasets [1]. Mazarka proposes OneFlow-a flow-based one-class classifier for anomaly detection that outperforms related methods on real-world anomaly detection problems [2]. Wentao presents a two-step outlier filtering framework for city-scale localization. Performance was shown on real-world datasets [3]. Chengzhi et al. propose CIMD - a new outlier removal approach that beats the state-of-the-art approaches very well [4]. The authors have presented a shade information-based method of outlier detection and a camera-projector correspondence method for improvement in outlier identification [5]. Soheil introduces the algorithm LOF for outlier detection. Owing to its satisfactory results in several accomplishments, this has been performed. Nurunnabi proposes robust statistical techniques towards outlier detection and robust saliency feature estimation. Considering this, accuracy and robustness have considerably improved. Authors have proposed a discriminative classifier for outlier detection in large-scale point clouds with high precision and accuracy. Based on the outlier factor of the cluster, a clustering-based outlier detection method, named CBOD, is presented, showing effectiveness and practicality [9]. The authors propose a novel method for outlier detection using a hierarchical spatial verification scheme, demonstrating its effectiveness and practicality [10]. It proposes an online unsupervised calibration algorithm depending only on a stationary target for automotive radars horizontally aligning accurately. Authors of [12] have presented a critical review, along with a categorization of

the outlier detection methods, concerning geomechanical data. Insights into fence labeling methods, statistical tests, and practical applications using real examples in this regard are provided. Yuan et al. proposed a new algorithm for data clustering called SNCRL, standing for Spatial Neighborhood Connected Region Labeling inspired by the spatial relationship which could help in outlier detection and handling in 3D point cloud data [13]. Kharroubi et al. reviewed the standard methods and recent advancements reached while using machine learning and deep learning in conducting change detection in point cloud data and gave a contribution to outlier detection techniques [14]. Abouelaziz et al. presented a framework to discuss various challenges related to the cleaning of building point clouds in outlier detection and removal, indicating some methodologies that may be applied to the treatment of outliers in 3D data [15]. Tychola et al., through a systematic literature review on several aspects of 3D point clouds, may reveal machine learning methods that could enhance the outlier detection when using standard deviation methods [16]. Jia et al. proposed a method of outlier detection and removal based on a dynamic standard deviation threshold and pointed out an approach that could be complementary to handle outliers using machine learning-enhanced standard deviation methods [17].

III. Research Environment

This study endeavors to identify and eliminate the outliers on real-world data, The data were collected using laser devices from inside the container, as depicted in Figure 1. The measurement occasionally incorporates dust particles and indiscernible values originating from both the location and the measurement process.

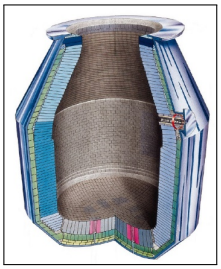


Fig. 1: A container to be measured by laser

In Figure 2, measurements are depicted for two distinct samples conducted at varying process times, each presented in a column. For each sample, the initial two rows display measurements from different angles, while the third row showcases the container divided into four parts and subsequently reassembled. Finally, the last row demonstrates measurements conducted in two dimensions.

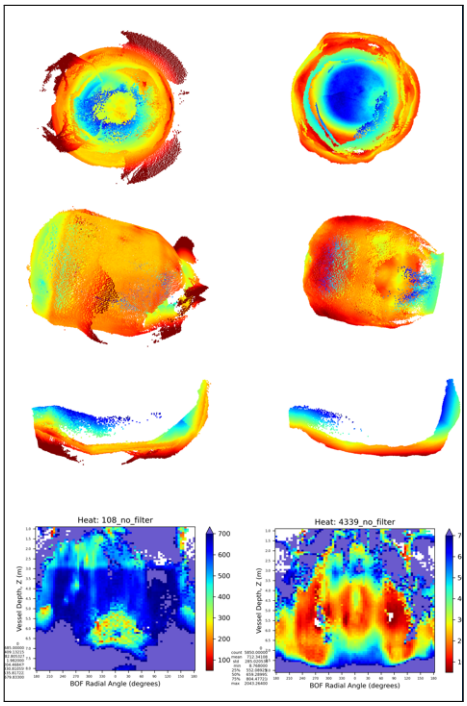


Fig. 2: Shows the measurements of a container in different process times from different angles

IV. Different Outlier Detection methods in 3D Measurements

In this paper, various methods and algorithms have been implemented and are listed in Table I along with their performance and outcomes. The successfully implemented ones, alongside their comparative results, are assessed in subsequent sections. The table is categorized into four sections: those unsuitable for this data type, ones with low performance and high execution time, algorithms yielding results far from the target, and finally, methods for potential evaluation. The outcomes from the latter categories reveal two primary challenges. Firstly, some methods may eliminate informative data, resulting in gaps. Secondly, another challenge arises when methods fail to detect outliers, erroneously treating them as inliers.

The following provides a concise summary of the comparative methods in Table I:

A. *Kernel Density Estimation (KDE) [18]*

It does this by searching for areas of low density in the probability distribution estimate. In many types of outliers, the defining characteristic is that they lie in regions where the density is substantially lower than the density that defines the majority of the data.

B. *K-Nearest Neighbors (KNN) [19]*

KNN is one of those versatile algorithms which are widely applied across domains for outlier detection. This algorithm works on the very basics of distance measures between the data points, which normally determines whether an observation is an outlier or not. KNN is a non-parametric, supervised

Methods Name	Performance Result
Elliptical Envelope	Comparative
Local Outlier Factor	
K Nearest Neighbors	
Isolation Forest	
Feature Bagging	
Empirical Cumulative Distribution Function	
Combination of cluster_dbscan and remove_sta	
Kernel Density Estimation	
Standard Deviation	
Unsupervised Outlier Detection Using Empirical	Non Comparative
Fast Angle-Based Outlier Detection	
Copula-Based Outlier Detection	
Principal Component Analysis	
Minimum Covariance Determinant	
Clustering-Based Local Outlier Factor	
Histogram-based Outlier Score	
Rotation-based Outlier Detection	
Single-Objective Generative Adversarial Actions	Low Performance
Multiple-Objective Generative Adversarial Actions	
Angle-Based Outlier Detection	
Connectivity-Based Outlier Factor	
Variational AutoEncoder	Improper for 3D point cloud
Deep One-Class Classification	
Locally Selective Combination of Parallel Outliers	
Accelerating Large-scale Unsupervised Heterogeneity	
Fully connected AutoEncoder	
Extreme Boosting Based	
Lightweight On-line	
Median Absolute Deviation	
One-Class Support Vector Machines	
Deviation-Based Outlier Detection	
Subspace Outlier Detection	
Fast Local Correlation Integral	

TABLE I: Algorithm Result Table

machine learning algorithm that can be adapted for unsupervised anomaly detection.

C. Local Outlier Factor (LOF) [20]

LOF is one among the popular outlier detection algorithms that works on a basic idea of local density. This algorithm studies the anomaly score of a data point w.r.t. its local density compared with its neighbors. Thus, LOF selects outliers with a lower local density than that of their neighbors, hence making it robust against data density variations.

D. Standard Deviation (SD)

The process involves the calculation of mean and standard deviation of the dataset. A data point that lies outside a threshold, normally multiple of standard deviation such as ±3 standard deviations is considered an outlier. That approach assumes normal distribution. This is most likely meant for the Density-Based Spatial Clustering of Applications with Noise, or DBSCAN. DBSCAN is a clustering algorithm which groups points that are proximal with each other in terms of density. This is quite often applied in the application of clustering in spatial data.

E. Remove Statistical Outlier [21]

It is intended for removing statistical outliers from the point cloud. The statistical outlier removal is a method for finding those points in the point cloud that are much far away from the mean concerning statistical measures. This helps in cleaning up noisy or inaccurate data in the point cloud.

F. Empirical Cumulative Distribution-based (ECOD) [22]

ECOD is a probability-based unsupervised outlier detection method dependent on empirical cumulative distribution functions. It detects outliers in the tail of a distribution by estimating and multiplying the empirical cumulative distribution functions of data points, hence making the approach interpretive and parameter-free, effective for outlier detection tasks.

G. Elliptical Envelope [20]

The Elliptical Envelope is an unsupervised anomaly detection method via machine learning. It works in a way that it would create an imaginary elliptical area around a dataset, assuming the latter is in a Gaussian distribution. It is really effective in outlier detection in datasets showing multivariate Gaussian distribution patterns, thus enabling the identification of considerably deviated data from the expected distribution. This capability makes it particularly useful in various anomaly detection scenarios.

H. Feature Bagging [23]

An ensemble-based outlier detection algorithm for very large, high-dimensional, and noisy databases, this represents an approach in which multiple algorithms of outlier detection run on a database with different sets of features. This algorithm provides an improved overall performance of outlier detection by combining the outputs of various algorithms.

I. Isolation Forest [24]

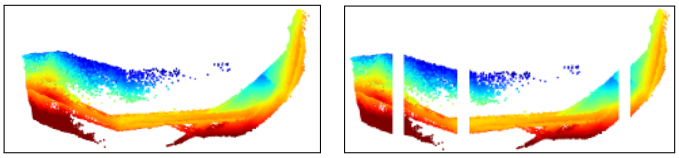
It is one of the anomaly detection systems that works on finding the abnormalities within a dataset with quite great speed. It is an unsupervised learning algorithm based on the binary tree concept. This algorithm labels anomalies recursively by partitioning the outliers.

V. ML-Enhanced Standard Deviation Methodology

Our proposed has been introduced as ML-Enhanced Standard Deviation. In order to solve the problem and introduce a better method to remove and detect the outliers in 3D environments, this solution uses Machine Learning. Furthermore, this approach uses the Cylindrical Coordinate System as its ground. Then each 3D point cloud measurement will cover Radius (R), Angle (A), Height (H), and Measurement (M). It has been implemented in following consecutive steps:

A.

KNN-clustering algorithm is being used to divide the point-cloud measurements into clusters in the early stage of our method implementation. In our approach, the point-cloud data is divided into four distinguished clusters. This division is shown in Figure 3.



(a) (b)

Fig. 3: Two smaller images placed horizontally

B.

So that the following steps can be carried out, for the realization of the methodology under proposal, we present a set of formulas from computational mathematics. These selected and designed-with-care formulas constitute the base of such a representation-something to be done in a methodical and rigorous way. In this direction, the application of these formulas is made to the variable of measurement, M. Let measurement $= \{x_1, x_2, \ldots, x_n\}$ be the set of values.

1. Cutoff Coefficient:

$$\text{CutoffCoefficient} = \frac{\text{mean}_{\{x_i\}}}{\text{Quantile}_{0.25}(\{x_i\})}$$

2. Cutoff:

$$\text{Cutoff} = \text{STD}_{\{x_i\}} \times \text{CutoffCoefficient}$$

3. Lower Inlier Bound:

$$\text{LowerInlier} = \text{mean}_{\{x_i\}} - \text{Cutoff}$$

4. Upper Inlier Bound:

$$\text{UpperInlier} = \text{mean}_{\{x_i\}} + \text{Cutoff}$$

5. Inlier Condition:

$$\text{inlier measurement} = \{x_i \mid x_i \geq \text{LowerInlier}$$
$$\& \; x_i \leq \text{UpperInlier}\}$$

6. Inlier Distance:

$$\text{InlierDistance} = \text{STD}_{\{x_i\}}$$

" Eqs. 3, 4, and 5 introduce an improved standard deviation method that should work by means of dynamic cutoff values presented in Eq. 1 and Eq. 2. It allows setting a variable cutoff threshold for every case.

C.

The output of Equation 5 represents the absolute and stringent inlier set, which is then utilized as labeled data to train a machine learning model. Figure 4 depicts the output of Equation 5 representing the inlier readings after the implementation of the algorithm.

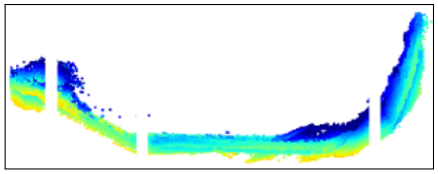


Fig. 4: Depicts the inliers which will be used for ML

D.

Thirdly, an extracted inlier labelled data based on height and angle features, and radius as a target, is used for training the machine learning algorithm. Notably, measurement is excluded in this context.

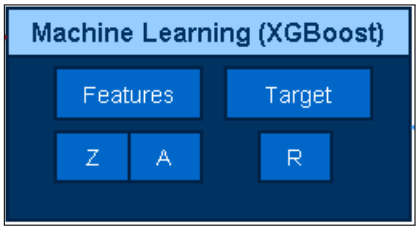
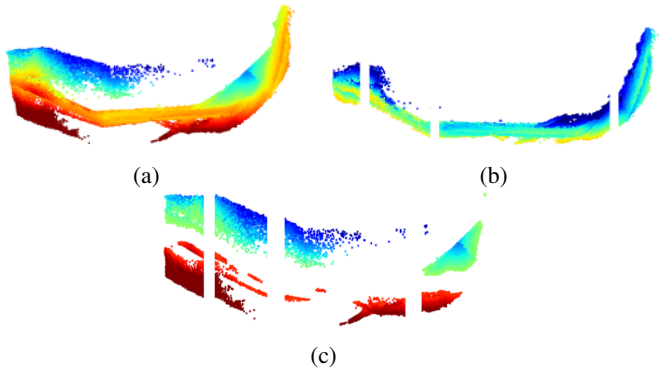


Fig. 5: Shows ML's inputs and models

E.

The suspected outliers are obtained by subtracting the raw data by the inlier labeled data of step 3.



(a) (b)

(c)

Fig. 6: The plot after subtracting the suspected outlier values

F.

The suspected outliers in previous step pass through the trained machine learning model to find the predicted radiuses.

G.

If the difference between the actual and predicted radius is less than the calculated thresholds of inlier distance in Eq. 6 then this measurement is labelled as an inlier, otherwise, it is labelled as an outlier.

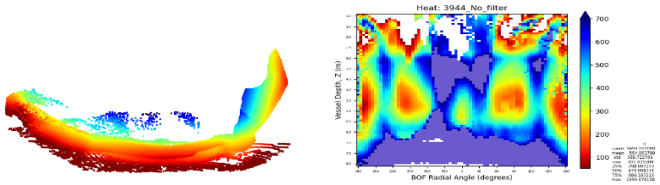
H.

The aggregate of pure inlier measurements in step 3 and the inlier result of step 7 are our total desired result

VI. Evaluation Results

The following plots showcase the outcomes for each methodology. Figure 7 presents the measurements without the application of any algorithms or methods, serving as reference data.

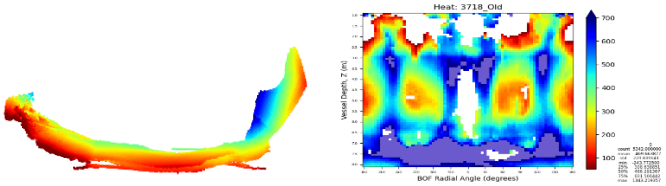
A. *Raw Measurement without Applying any Filter*



(a) (b)

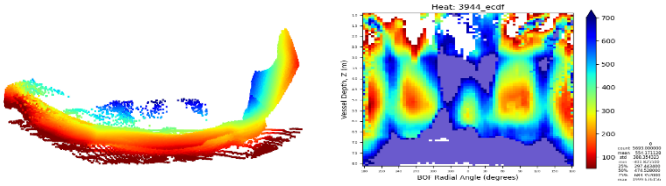
Fig. 7: Raw Measurement

In the fourth section the successfully implemented algorithms, with comparative results have been applied on the raw measurement data, the figures 8 to 16 depict these outcomes.



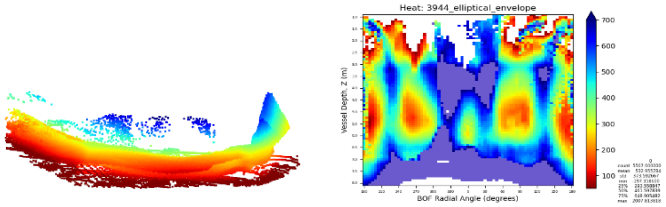
(a) (b)

Fig. 8: O3D



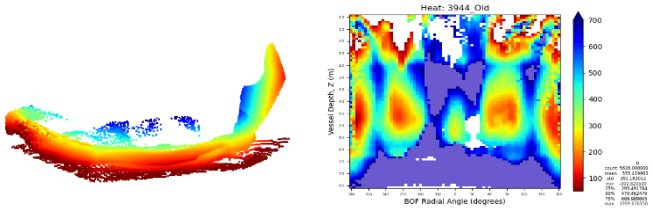
(a) (b)

Fig. 9: ECDF



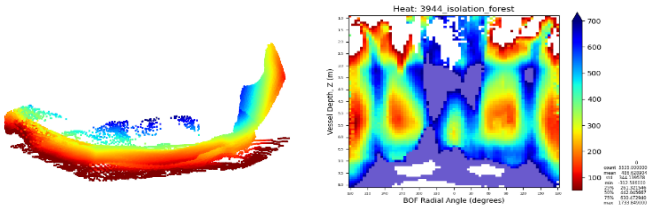
(a) (b)

Fig. 10: Elliptical Envelope



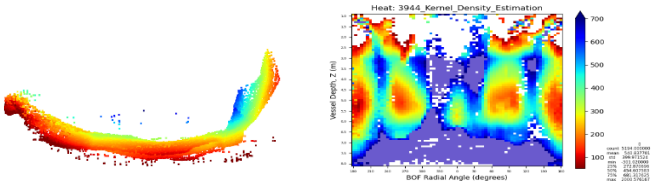
(a) (b)

Fig. 11: Feature Bagging



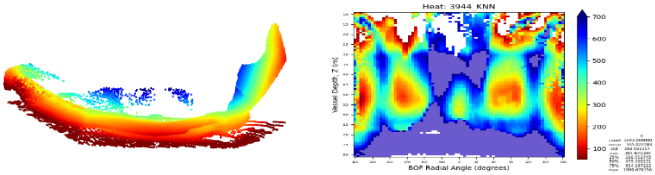
(a) (b)

Fig. 12: Isolation Forest



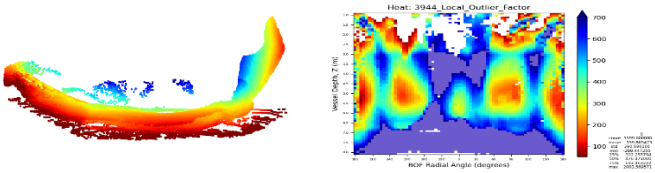
(a) (b)

Fig. 13: Kernel Density Estimation



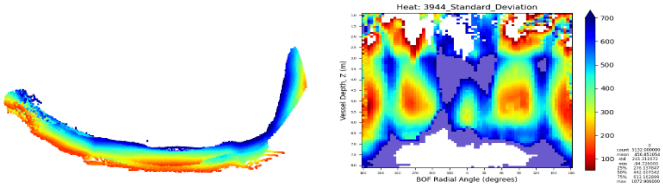
(a) (b)

Fig. 14: K-Nearest Neighbors



(a) (b)

Fig. 15: Local Outlier Factor

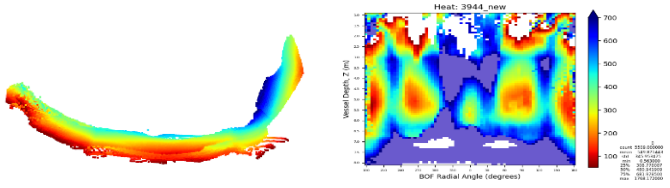


(a) (b)

Fig. 16: Standard Deviation

It is evident that the aforementioned methods may inadvertently eliminate informative data, or in some cases, they may fail to effectively identify and remove genuine outliers. Figure 17 depicts the ML-Enhanced Standard Deviation method, The analysis demonstrates a remarkable outcome wherein not only

was informative data retained, but genuine outliers were also successfully eliminated.



(a) (b)

Fig. 17: ML-Enhanced Standard Deviation

Figure 18: Comparative analysis of all methods based on inlier data count. It is observed that when comparing the new method or ML-Enhanced Standard Deviation to other vicinity algorithms, the new method returns very remarkable results. Unlike the rest, it has overcome various challenges such as removing too many inliers that may create gaps in measurements, or it retains too many outliers within the dataset.

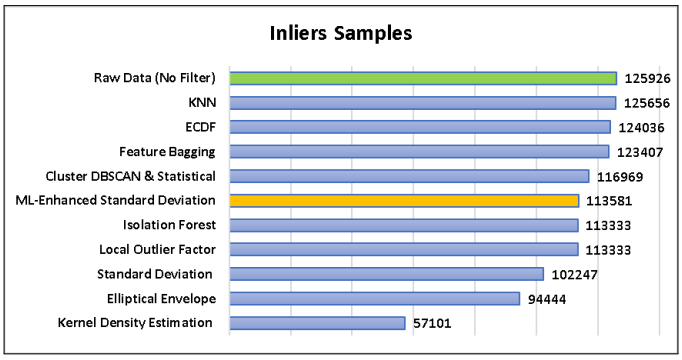


Fig. 18: Inlier dataset size after applying different algorithms

VII. Conclusion

In this paper, a new two-step strategy for outlier filtering using real inliers is proposed. It employs an accurate statistical method to identify real inliers. This refined dataset then acts as an input for a model based on machine-learning techniques. The model classifies the validation of the detected measurements to either be inliers or outliers. Most importantly, our approach is embodied with new mathematical and statistical formulas. In fact, we proved the efficiency and effectiveness of our proposed outlier filtering framework through an intensive real-dataset-based evaluation.

References

[1] M. Yuan, Z. Li, K. W. Wan and W. Y. Yau, "Outlier Detection using Hierarchical Spatial Verification for Visual Place Recognition," 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), Singapore, 2018, pp. 1-6, doi: 10.1109/ICARCV.2018.8581070.

[2] Ł. Maziarka, M. Śmieja, M. Sendera, Ł. Struski, J. Tabor and P. Spurek, "OneFlow: One-Class Flow for Anomaly Detection Based on a Minimal Volume Region," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 8508-8519, 1 Nov. 2022, doi: 10.1109/TPAMI.2021.3108223.

[3] W. Cheng, K. Chen, W. Lin, M. Goesele, X. Zhang and Y. Zhang, "A Two-Stage Outlier Filtering Framework for City-Scale Localization Using 3D SfM Point Clouds," in IEEE Transactions on Image Processing, vol. 28, no. 10, pp. 4857-4869, Oct. 2019, doi: 10.1109/TIP.2019.2910662.

[4] C. Qu, Y. Zhang, K. Huang, S. Wang and Y. Yang, "Point Clouds Outlier Removal Method Based on Improved Mahalanobis and Completion," in IEEE Robotics and Automation Letters, vol. 8, no. 1, pp. 17-24, Jan. 2023, doi: 10.1109/LRA.2022.3221315.

[5] H. T. N. Dung and S. Lee, "Outlier removal based on boundary order and shade information in structured light 3D camera," 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), Siem Reap, Cambodia, 2015, pp. 124-129, doi: 10.1109/ICCIS.2015.7274608.

[6] Sotoodeh, Soheil. "OUTLIER DETECTION IN LASER SCANNER POINT CLOUDS." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2006): 297-302.

[7] Nurunnabi, Abdul Awal Md, Geoff A. W. West and David Belton. "Outlier detection and robust normal-curvature estimation in mobile laser scanning 3D point cloud data." Pattern Recognit. 48 (2015): 1404-1419.

[8] Stucker, Corinne, Audrey Richard, Jan Dirk Wegner and Konrad Schindler. "SUPERVISED OUTLIER DETECTION IN LARGE-SCALE MVS POINT CLOUDS FOR 3D CITY MODELING APPLICATIONS." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2018): n. pag.

[9] Jiang, Sheng-yi and Qing-bo An. "Clustering-Based Outlier Detection Method." 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery 2 (2008): 429-433.

[10] M. Yuan, Z. Li, K. W. Wan and W. Y. Yau, "Outlier Detection using Hierarchical Spatial Verification for Visual Place Recognition," 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), Singapore, 2018, pp. 1-6, doi: 10.1109/ICARCV.2018.8581070.

[11] A. Bobaru, C. Nafornita and V. C. Vesa, "Unsupervised Online Horizontal Misalignment Detection Algorithm for Automotive Radar," 2022 14th International Conference on Communications (COMM), Bucharest, Romania, 2022, pp. 1-5, doi: 10.1109/COMM54429.2022.9817178

[12] Dastjerdy, Behzad, Ali Saeidi, and Shahriyar Heidarzadeh. 2023. "Review of Applicable Outlier Detection Methods to Treat Geomechanical Data" Geotechnics 3, no. 2: 375-396. https://doi.org/10.3390/geotechnics3020022

[13] Yuan, Xiaocui, Huawei Chen and Bao Ling Liu. "Point cloud clustering and outlier detection based on spatial neighbor connected region labeling." Measurement and Control 54 (2020): 835 - 844.

[14] Kharroubi, Abderrazzaq, Florent Poux, Zouhair Ballouch, Rafika Hajji and Roland Billen. "Three Dimensional Change Detection Using Point Clouds: A Review." Geomatics (2022): n. pag.

[15] Abouelaziz, Ilyass and Youssef Mourchid. "A Framework for Building Point Cloud Cleaning, Plane Detection and Semantic Segmentation." ArXiv abs/2402.00692 (2024): n. pag.

[16] Tychola, Kyriaki A., Eleni Vrochidou and George A. Papakostas. "Deep learning based computer vision under the prism of 3D point clouds: a systematic review." The Visual Computer (2024): n. pag.

[17] Jia, Chaochuan, Ting Yang, Chuanjiang Wang, Binghui Fan and Fu Gui He. "A new fast filtering algorithm for a 3D point cloud based on RGB-D information." PLoS ONE 14 (2019): n. pag.

[18] Latecki, L.J., Lazarevic, A., Pokrajac, D. (2007). Outlier Detection with Kernel Density Functions. In: Perner, P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2007. Lecture Notes in Computer Science(), vol 4571. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73499-4-6

[19] T. T. Dang, H. Y. T. Ngan, and W. Liu, "Distance-based k-nearest neighbors outlier detection method in large-scale traffic data," 2015 IEEE International Conference on Digital Signal Processing (DSP), Singapore, 2015, pp. 507-510, doi: 10.1109//ICDSP.2015.7251924.

[20] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[21] https://www.open3d.org/docs/0.11.0/tutorial/geometry/pointcloud_outlier_removal.html.

[22] Li, Zheng, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu and George H. Chen. "ECOD: Unsupervised Outlier Detection Using Empir-

ical Cumulative Distribution Functions." IEEE Transactions on Knowledge and Data Engineering 35 (2022): 12181-12193.
[23] Lazarevic, Aleksandar and Vipin Kumar. "Feature bagging for outlier detection." Knowledge Discovery and Data Mining(2005).
[24] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation Forest."2008 Eighth IEEE International Conference on Data Mining(2008): 413-422.