Heterogeneous Point Clouds Matching using Supervoxel Signatures from a Deep Neural Network Autoencoder

Tsung-Han Wen,[1] Tee-Ann Teo*

Dept. of Civil Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan 30010
(edwen.en08@nycu.edu.tw, tateo@nycu.edu.tw)

Keywords: Lidar, Autoencoder, Deep Learning, Supervoxel, Matching, Data Compression.

Abstract

Advancements in lidar systems have improved the performance of 3D data acquisition. Differences arise between the point clouds obtained by different lidar sensors, such as variations in point density, random error, and scanning patterns. This study presents a novel approach for automatic cross-sensor matching of lidar point clouds using a deep neural network autoencoder (DNN-AE) and supervoxel signatures. A compact representation called a supervoxel signature was formed by voxelizing and reprojecting the point clouds, generating multiscale supervoxels, and encoding them with a DNN-AE. The proposed method demonstrated high matching accuracy and tolerance to point density differences and random registration, showcasing its effectiveness in addressing the challenges associated with varying lidar sensor data. From the simulation results, the supervoxel signature had a matching correctness of 83.78% when the point density was 1/256 of the original one, and the tolerance to random errors reached the submeter level. In addition, the multiscale supervoxel signature was more reliable than the single-scale combination. In real-world cross-sensor experiments involving consumer-grade and surveying-grade lidar systems, the proposed method achieved a matching accuracy exceeding 90% by aggregating features across adjacent frames, while significantly reducing data volume. These results confirm the robustness and practicality of the proposed framework for reliable and efficient heterogeneous point cloud matching.

1. Introduction

Mobile lidar systems (MLSs) have emerged as an efficient method for acquiring 3D point clouds in road environments. Advances in lidar technology have made 3D data collection faster and more accessible. However, different lidar sensors, such as the Riegl VMX250 for mapping and the Velodyne VLP16 for autonomous navigation, produce point clouds with varying densities, scanning patterns, and levels of accuracy. As a result, point clouds from different lidar systems often differ significantly in structure, making cross-sensor registration a complex and ongoing research challenge. Point cloud registration is crucial for integrating heterogeneous lidar datasets into a common spatial framework. This process estimates the geometric transformation between datasets to enable consistent analysis and mapping. However, variations in density, noise levels, and coverage areas complicate automatic alignment (Huang et al., 2021; Zhao et al., 2025). Most existing approaches rely on a coarse-to-fine strategy, where an initial coarse match guides subsequent refinement (Yang et al., 2024; Xu et al., 2025). Despite progress, automatic coarse registration across different lidar systems remains difficult due to the lack of robust and generalizable feature representations (Zhao et al., 2025). Previous studies have explored various coarse matching strategies to address this challenge, broadly categorized into point-based, line-based, surface-based, and deep learning-based methods.

Point-based methods such as Iterative Closest Point (ICP) (Besl and McKay, 1992) have been widely used for point cloud registration. However, ICP is often sensitive to noise and point density variations, and it typically requires a good initial alignment, which limits its effectiveness in cross-sensor applications (Xu et al., 2025). To improve robustness, keypoint-based strategies have been developed, using distinct features such as object centroids to guide alignment between different datasets (Nagy and Benedek, 2018). Line-based methods take advantage of the geometric stability of linear features, which tend to be more reliable than individual points. Common features include building outlines (Yang et al., 2015) and road edges (Javanmardi et al., 2017), which can be extracted using statistical models and matched using probabilistic alignment techniques like the Normal Distribution Transform (NDT) (Javanmardi et al., 2017). Surface-based methods extract geometric information from planar or curved regions, making them more robust to noise than point- or line-based approaches. Lidar systems can capture a wide range of surface features, such as ground surfaces, building roofs, and walls, which are particularly useful for accurate cross-sensor registration. These methods typically rely on surface normals and structural consistency to align point clouds from different sensors (Zhang et al., 2012; Wu et al., 2014; Teo and Huang, 2014).

Recent advances in deep learning have significantly enhanced 3D point cloud registration, particularly through learned feature extraction using neural networks. Early approaches, such as Elbaz et al. (2017), divided point clouds into super points using a random sphere cover algorithm, projected them onto depth maps, and applied an autoencoder (AE) for feature compression. Similarity was then evaluated using Euclidean distance, and the top transformation candidates were refined via RANSAC. Huang (2017) proposed a method based on deep convolutional neural networks (CNNs), using structured 3D data transformed through the truncated distance function to produce high-dimensional descriptors for cross-sensor matching. Beyond these foundational works, newer architectures have leveraged Transformer models for more expressive 3D representations. For instance, Point-BERT (Yu et al., 2022) adapted the BERT framework to point clouds using a masked point modeling strategy and discrete variational autoencoders (dVAE) for semantic tokenization. Fu

* Corresponding author.

et al. (2023) improved this design by introducing multi-choice tokens to better handle ambiguity in point cloud encoding, thereby enhancing performance in classification and registration tasks.

With the increasing variety of sensors, recent studies have emphasized registration methods for heterogeneous point clouds to effectively handle discrepancies arising from diverse sensor sources in large-scale outdoor environments. Jia et al. (2024) introduced an incremental registration method using hierarchical graph matching, which constructs multiscale graphs of source and target scans and performs coarse-to-fine registration by matching them with refined structural and feature constraints. Xu et al. (2025) proposed a multi-source fine-registration strategy combining fully-connected feature graphs, heat conduction–based correspondence selection, and weighted least-squares optimization, achieving high-precision alignment even with inconsistent hardware and noisy data. Zhao et al. (2025) proposed Cross-PCR, a framework for heterogeneous point cloud registration that combines local geometry descriptors, overlap prediction, and confidence-guided matching to enhance correspondence reliability and transformation accuracy.

Place recognition is particularly essential in heterogeneous point cloud scenarios, where differences in point density, scanning patterns, and accuracy complicate accurate localization. It facilitates coarse localization by identifying whether a current scan corresponds to a previously observed location, significantly simplifying subsequent point cloud matching. Traditional place recognition methods often employ global descriptors designed to handle significant viewpoint variations and environmental changes. For example, Scan Context (Kim and Kim, 2018) leverages a polar-coordinate-based descriptor to represent spatial distributions of points, enabling rapid and robust recognition. More recent approaches incorporate deep learning to enhance discriminative power and invariance to environmental conditions. PointNetVLAD (Uy and Lee, 2018) and its successor, LPD-Net (Liu et al., 2019), integrate point cloud feature extraction via neural networks with VLAD-based aggregation, demonstrating superior performance under varied sensor setups and challenging environmental scenarios. Zou et al. (2023) proposed PatchAugNet, introducing a patch-wise feature augmentation strategy that injects randomness at the feature level to simulate variations across heterogeneous point clouds and enhance local feature robustness. Xu et al. (2022) proposed a lightweight framework for heterogeneous place recognition, using a virtual LiDAR simulation module to bridge domain gaps and polar grid height coding to provide compact, rotation-invariant representations.

Simplifying raw point clouds while preserving geometric semantics is also critical for robust cross-source matching (Jia et al., 2024). Moreover, directly processing unordered points with models like PointNet often leads to increased model complexity and high computational costs, limiting deployment in practical applications (Komorowski, 2021; Hui et al., 2022). To address challenges such as irregular point distribution and computational inefficiency, voxel-based representations have become increasingly popular for point cloud processing. Voxels enhance computational efficiency and feature extraction by organizing raw 3D points into structured grids. Xu et al. (2021) leveraged voxel-based surface constraints for automated coarse registration, while Li et al. (2022a) used adaptive region growing to extract planar features within voxel grids. Xiong et al. (2021) introduced density gradient simplification to accelerate keypoint detection, and Li et al. (2022b) demonstrated that voxelized data supports fast, parallel registration through their lightweight VPRNet framework.

Building on these concepts, this study proposes a streamlined and effective alternative with a local descriptor design for heterogeneous point cloud matching: *Supervoxel Signatures*, a compact multiscale feature representation learned via deep neural network autoencoders (DNN-AE). The features of the proposed supervoxel signature are simple. Specifically, we focused on low-density lidar (LDLidar) and high-density lidar (HDLidar) point clouds acquired from different lidar systems. In this study, LDLidar refers to a consumer-grade advanced driver assistance system lidar sensor (i.e., Velodyne), while HDLidar refers to a surveying-grade lidar sensor (i.e., Riegl) for high-density mapping purposes. In practical operations, surveying-grade HDLidar point clouds are geocoded into the world coordinate system. To register an automotive-grade LDLidar as an HDLidar, it needs to be transformed into the world coordinate system. It can be used to improve the positioning accuracy of the automotive-grade LDLidar. Both point clouds were first voxelized to support effective feature extraction, and a DNN-AE was then applied for feature learning. The resulting features, called *supervoxels*, were compressed and stacked to form a supervoxel signature for cross-sensor matching. This approach proved to be an effective data compression method, significantly reducing the volume of 3D data and improving the efficiency of place recognition for coarse matching.

The main contribution of this study lies in structurally addressing three core challenges in heterogeneous point cloud matching—density variation, noise, and rotation—while simultaneously achieving efficient compression and reliable matching, all without relying on complex network architectures.

2. Methodology

This study proposes a feature extraction framework for heterogeneous lidar point clouds to generate consistent codes across different systems. The method extracts reliable supervoxel signatures from large sets of 3D points, serving as the basic unit for coarse matching between HDLidar and LDLidar systems. The workflow consists of three modules: (1) the Structuralization module, (2) the Encoder module, and (3) the Matching module. As illustrated in Figure 1, the target points refer to georeferenced HDLidar data, while the sensed points denote LDLidar data requiring alignment.

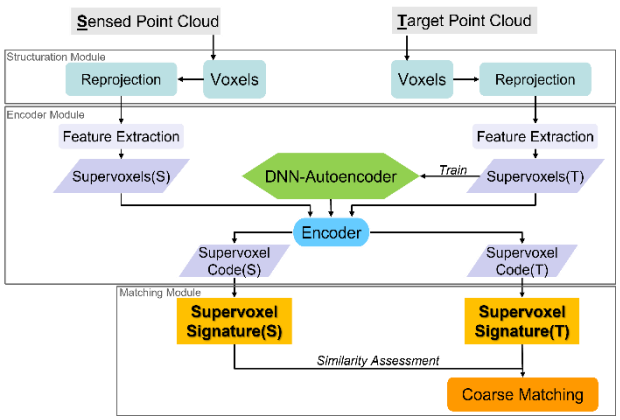

Figure 1. Workflow of the proposed scheme.

2.1 Structuralization Module

Irregular point clouds are unordered sets of 3D vectors, making point-wise processing computationally intensive, particularly at large scales. This unstructured nature limits the efficiency of direct analysis. Voxelization addresses the issue by grouping nearby points into regular grids, reducing redundancy and enabling more efficient, segmentation-like processing. In this study, we partitioned the area of interest into a voxel grid using a maximum voxel size (V_{max}=10 m). Each point was assigned to a voxel based on its spatial index, as defined in Equation (1). To capture multiscale features, the grid was subdivided into finer levels (e.g., V_2=5 m, V_3=2.5 m). Since our data primarily consists of MLS scans in road environments, voxelization was restricted to the XY plane to reduce the computational cost, effectively treating the process as 2D rasterization of 3D point clouds. Figure 2 illustrates the multiscale voxelization concept. After voxelizing each scale, we calculated the normal vector for each voxel using all points within the V_{max} voxel. This normal was then used to reproject the point cloud, minimizing the effects of varying scan angles. Principal Component Analysis (PCA) was applied to the covariance matrix of each point set to derive eigenvalues and eigenvectors. The eigenvector corresponding to the largest eigenvalue is defined as the principal axis. Using the three eigenvectors as basis vectors, we rotated the point cloud to align with the principal axis.

$$\begin{cases} i_p = int\left(\frac{x_p - x_0}{\Delta x}\right) \\ j_p = int\left(\frac{y_p - y_0}{\Delta y}\right), \\ k_p = int\left(\frac{z_p - z_0}{\Delta z}\right) \end{cases} \quad (1)$$

where (i_p, j_p, k_p) is the node index of voxel; (x_p, y_p, z_p) is the 3D lidar point; (x_0, y_0, z_0) is the origin of the coordinates; and $(\Delta_x, \Delta_y, \Delta_z)$ is the voxel size.

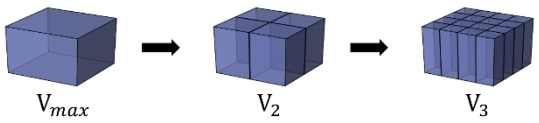

Figure 2. Illustration of multiscale voxelization.

2.2 Encoder Module

A voxel stores lidar points within a predefined 3D space, while a supervoxel is a cluster of features derived from those points. In this study, we transformed each voxel into a supervoxel by extracting intensity and geometric properties from its constituent points. Intensity features were calculated as the mean and standard deviation of the lidar intensity values, while geometric features were derived from the eigenvalues of the covariance matrix of the points within the voxel. As shown in Table 1, these geometric features were computed using various combinations of the eigenvalues ($\lambda_1 \geq \lambda_2 \geq \lambda_3$). In total, ten features were extracted per supervoxel and used as input to the DNN-AE. To capture multiscale characteristics, we applied voxelization at multiple resolutions. Figure 3 shows one of the ten features across different voxel sizes within the same region, where (a) represents the largest voxel size (V_{max}), and V_2, V_3, and V_4 are recursively half the size of the preceding scale.

Regarding data compression performance, a DNN-AE outperforms traditional PCA, particularly when appropriate compact dimensions are utilized. This is because a DNN-AE employs multiple hidden layers, which enhance its compression capability, surpassing that of PCA. The architecture of a DNN-AE comprises an encoder and a decoder section. The encoder section involves dimensionality reduction, beginning with the input layer and concluding with the bottleneck layer, via several fully connected hidden layers. The decoder section is the opposite of the encoder section, starting with the compact representation from the bottleneck layer and ending with the output layer. Additionally, the output dimensions should match the input dimensions based on the AE architecture, enabling the loss function to be defined effortlessly between the input and output layers. The key feature of a DNN-AE is its ability to utilize a deep neural network for learning processes within both the encoder and decoder sections while optimizing the network by minimizing the loss function.

In this study, we created multiscale supervoxels by combining different voxels of varying sizes. We then trained the compressed representation of the supervoxels, called the *supervoxel code*, using a DNN-AE approach. Each supervoxel scale required DNN-AE training in the proposed network's design. After extensive empirical testing and adjustments, we introduced a flexible and compact dimension network (see Figure 4), comprising eight fully connected hidden layers, with some layers that used ReLU as activation functions. We applied batch normalization before the activation function to reduce the training time and mitigate the vanishing gradients problem. The symbol C_s denotes the adjustable compact dimension, the supervoxel code size, while I_s and O_s represent the input and output sizes, respectively. In order to train the autoencoder model, we employed mean square error as the loss function and opted for the Adam optimizer. After completing the training of the DNN-AE, we could use the encoder section for data compression. We utilized the supervoxel code as the compressed feature, which was generated by the supervoxels through the DNN-AE encoder.

Feature	Formula
Curvature	$C_\lambda = \lambda_3/(\lambda_1 + \lambda_2 + \lambda_3)$
Linearity	$L_\lambda = (\lambda_1 - \lambda_2)/\lambda_1$
Planarity	$P_\lambda = (\lambda_2 - \lambda_3)/\lambda_1$
Scattering	$S_\lambda = \lambda_3/\lambda_1$
Omnivariance	$O_\lambda = (\lambda_3 * \lambda_2 * \lambda_1)^{1/3}$
Anisotropy	$A_\lambda = (\lambda_1 - \lambda_3)/\lambda_1$
Eigenentropy	$E_\lambda = \lambda_1 \ln(\lambda_1) + \lambda_2 \ln(\lambda_2) + \lambda_3 \ln(\lambda_3)$
Eigensum	$\Sigma_\lambda = \lambda_1 + \lambda_2 + \lambda_3$

Table 1. Geometric features based on the eigenvalues

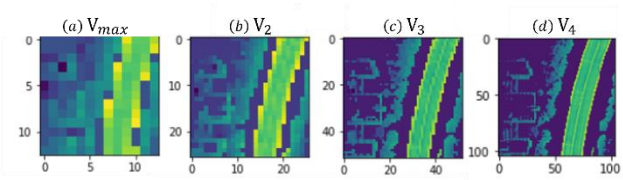

Figure 3. Illustration of a multiscale supervoxel.

https://doi.org/10.5194/isprs-annals-X-4-W6-2025-217-2025

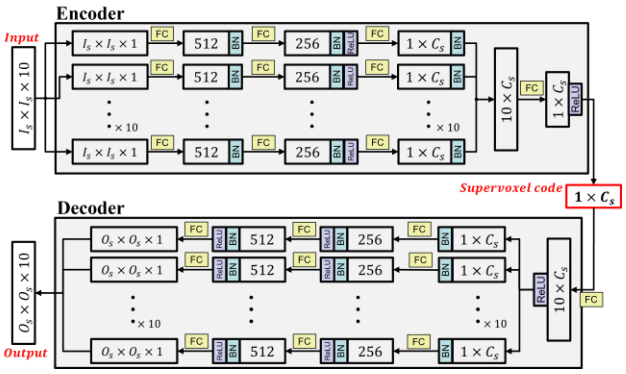

Figure 4. Proposed DNN-AE architecture.

2.3 Matching Module

To obtain multiscale supervoxels, we structuralized the point clouds in the same region with different voxel sizes. Then, we performed DNN-AE encoding on each supervoxel separately to acquire the supervoxel code at their corresponding scales. This code was the compressed representation of the supervoxel itself and could be used as the basic unit for the matching process. We stacked each scale's supervoxel code to produce a deeper code called a *supervoxel signature*. The supervoxel signatures were built up from the supervoxel code at different scales. The multiscale combination of the supervoxel signatures enabled us to describe the characteristics of the point clouds at different scales, leading to better classification accuracy under challenging conditions. The idea of the supervoxel signature was similar to the keypoint descriptor for 2D image matching. To gain multiscale features, we designed three types of supervoxel signatures based on scale combination: (1) single-scale, (2) dual-scale, and (3) multiscale (see Table 2).

Single	$V_{max} = 10m$	$V_2 = 5m$	$V_3 = 2.5m$	$V_4 = 1.25m$
Dual	$V_{max} + V_2$		$V_2 + V_3$	$V_3 + V_4$
Multi-	$V_{max} + V_2 + V_3$		$V_2 + V_3 + V_4$	$V_{max} + V_2 + V_3 + V_4$

Table 2. Scale combination of the supervoxel signatures

The supervoxel signature served as a similarity index to determine the similarity between two sets of point clouds. Each point cloud data in the datasets was encoded separately through the aforementioned procedures. Therefore, the matching based on similar supervoxel signatures could be extended to the application of coarse matching. The two sets of overlapped lidar point clouds were divided into many segments, and each segment was processed into a supervoxel through voxelization and feature extraction. Next, the supervoxel code was encoded using the encoder of a DNN-AE and concatenated as a supervoxel signature. Similar supervoxel signatures in the two datasets were clustered to fulfill the need for cross-sensor matching.

A supervoxel signature should exhibit scale-invariant, independent, and reliable properties to facilitate matching. Consequently, different lidar systems that scan the same area should yield similar encoded features. This study defined cross-sensor coarse matching as the similarity assessment between the supervoxel signatures obtained from LDLidar and HDLidar. As indicated in Equations (2), the Euclidean distance was employed to evaluate the similarity.

$$\text{Euclidean distance} = |u - v| \quad (2)$$

where, u and v are flattened supervoxel signature vectors from LDLidar and HDLidar, respectively.

To enhance the robustness of signature-based matching, we further developed a sequence-to-sequence matching and geometric verification pipeline. Each sequence of supervoxel signatures from the sensed point cloud was compared against the target sequence using a sliding-window strategy similar to SeqSLAM (Milford and Wyeth, 2012). For each windowed tracklet, we computed the mean Euclidean distance over time-aligned segments and selected the best-matching offset with the minimal average distance. This approach exploits local spatial continuity to improve matching reliability for practical applications.

Following sequence matching, one possible extension is to extract the centroid coordinates of the matched segments and apply RANSAC-based rigid transformation estimation to eliminate outliers. This geometric verification step estimates the optimal rotation and translation using inlier pairs, typically defined based on spatial proximity within a threshold. Such integration of appearance-based and geometry-based constraints can further improve the accuracy and robustness of heterogeneous point cloud matching. However, this aspect is beyond the scope of the present study and is not further explored here.

3. Experimental Results

This study presents a lightweight and practical deep learning framework for heterogeneous point cloud matching. The utilization of the supervoxel signature, encoded via a DNN-AE, mainly facilitated cross-sensor matching. Each supervoxel signature encapsulated features extracted from the point clouds. Simulation experiments were conducted to better understand the function of supervoxel signatures in distinguishing point clouds captured by different lidar sensors. After confirming the efficacy of the supervoxel signatures for matching, subsequent analysis was conducted using real datasets obtained from Velodyne and Rigel lidars.

3.1 Simulation Analysis

To ensure effective cross-sensor registration, a supervoxel signature must satisfy three key properties: *scale-invariance*, which ensures consistent feature representation across varying point densities; *independence*, which enables robust encoding even in repetitive environments like urban road networks; and *reliability*, which maintains stable performance under noise and sensor-related variation. To validate these properties, we conducted simulation experiments using two overlapping HDLidar datasets acquired along different trajectories on an expressway in Taipei City, captured by a Riegl VMX250 MLS system. The original HDLidar data served as the target, while the sensed data (LDLidar) were simulated by resampling, rotating, applying drift, and adding random noise. Both datasets were divided into 37 frames of 130 m × 130 m with 50% overlap (see Figure 5). A series of four tests were performed to evaluate the robustness of the supervoxel signatures: (1) density, (2) rotation, (3) drift, and (4) noise test.

In this simulation-based evaluation, matching accuracy was assessed by calculating the Euclidean distance between supervoxel signatures on a per-frame basis. Specifically, for each sensed frame, its supervoxel signature was compared with that of all target frames in the feature space, and the one with the minimum distance was considered the best match. This frame-

level signature comparison approach allowed us to quantify the robustness and discriminative power of the supervoxel signature.

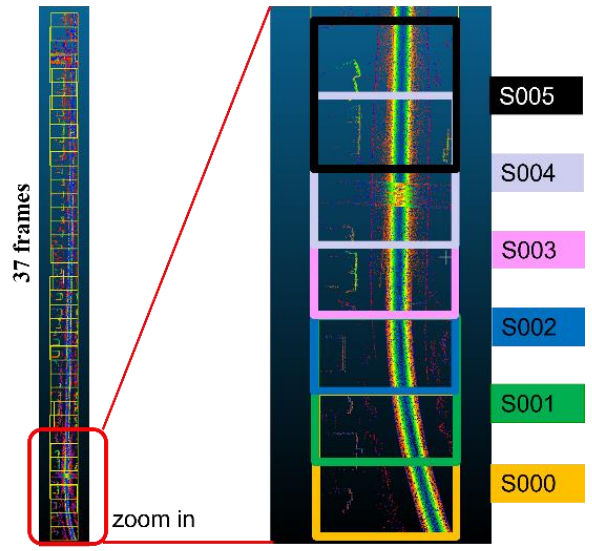

Figure 5. Simulated sensed point clouds (LDLidar).

3.1.1 Density Test: HDLidar and LDLidar datasets typically differ in point density. To evaluate the robustness of supervoxel signatures under varying densities, we simulated different density levels by randomly subsampling the original sensed point cloud (347.96 pts/m²), while the target point cloud remained at its original density (712.91 pts/m²). Eight density levels, from 1/2 to 1/256, were generated in powers of two (2^{-1}~2^{-8}). Each subsampled dataset was encoded using the same pre-trained encoder from the target data, and the resulting supervoxel signatures were used for matching based on Euclidean distance. In this analysis, the supervoxel code size was fixed at 16 to maintain consistent compression performance, while the voxel size was varied to assess feature extraction at different spatial resolutions. As shown in Figure 6, multiscale combinations generally outperformed single-scale configurations. The dual-scale combination V_{max}+V_2 achieved the highest accuracy, maintaining 83.78% even at 1/256 of the original density. In contrast, combinations involving only finer voxel sizes (e.g., V_3, V_4) were more sensitive to reduced point density and resulted in lower accuracy. These results demonstrate the effectiveness of multiscale supervoxel signatures, particularly those anchored by V_{max}, in handling large variations in point cloud density.

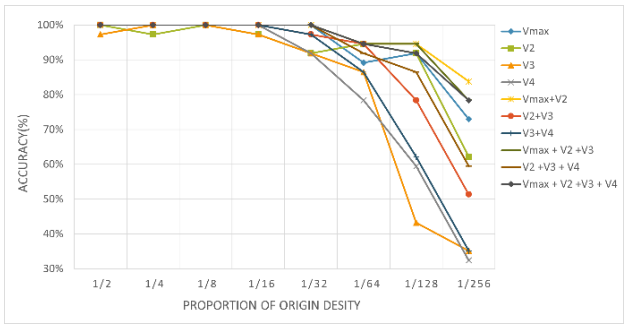

Figure 6. Matching accuracy of the density test.

3.1.2 Rotation Test: Most MLSs are equipped with positioning and orientation systems (POSs), such as GNSS, INS, and distance measuring instruments, to record location and attitude information during scanning. However, the accuracy of heading information varies across POS technologies. To evaluate the rotational sensitivity of supervoxel signatures, we simulated heading rotations ranging from $-15°$ to $15°$, using subsampled sensed point clouds at 1/32 of the original density to increase the challenge of the test. The target point cloud remained fixed, while the sensed point cloud was rotated around the center of each supervoxel, generating 31 rotated datasets, which were then encoded for matching. Figure 7 shows that single-scale signatures using small voxel sizes (e.g., V_3, V_4) and their combinations (e.g., V_3+V_4) were highly sensitive to rotation, resulting in low matching accuracy. In contrast, larger voxel sizes (V_{max}, V_2) and multiscale combinations (V_{max}+V_2+V_3, V_{max}+V_2+V_3+V_4) exhibited greater rotation tolerance, maintaining over 80% accuracy within $\pm 8°$ of rotation. This level of robustness aligns with typical POS heading accuracy (3°~4°), indicating that supervoxel signatures are well-suited for MLS applications involving real-world orientation variations.

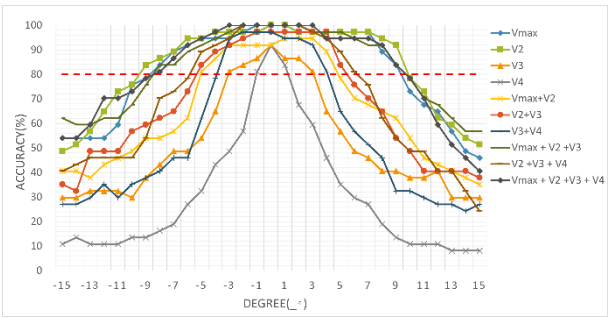

Figure 7. Matching accuracy of the rotation test.

3.1.3 Drift Test: Following the rotation test, a drift sensitivity test was conducted using sensed point clouds subsampled to 1/32 of the original density. This simulation aimed to assess the robustness of supervoxel signatures against trajectory drift, modeled as horizontal displacement ranging from 1 m to 10 m. The target point cloud remained fixed, while the sensed data were shifted horizontally to simulate positional drift in MLS systems. As shown in Figure 8, multiscale combinations such as V_{max}+V_2 and V_{max}+V_2+V_3 exhibited strong tolerance to drift, maintaining over 80% matching accuracy even with 4 m of displacement, approximately the width of a traffic lane. In contrast, single-scale signatures using small voxel sizes (e.g., V_3, V_4) and their combinations (e.g., V_3+V_4) were more sensitive to drift, showing a noticeable drop in accuracy. These results indicate that multiscale supervoxel signatures are more resilient to horizontal shifts and suitable for MLS scenarios with moderate positioning uncertainty.

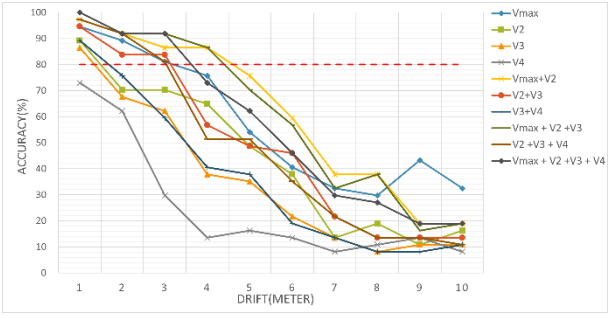

Figure 8. Matching accuracy of the drift test.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W6-2025-217-2025 | © Author(s) 2025. CC BY 4.0 License.

3.1.4 Tolerance of Noise: Laser ranging precision, observation direction, and beam divergence can affect lidar measurements' accuracy. In particular, differences between near and far scanning ranges and variations in the laser footprint may introduce random errors. To evaluate the robustness of supervoxel signatures to such noise, we added random errors to each point's x, y, and z coordinates in a subsampled sensed point cloud. A predefined standard deviation controlled the magnitude of the noise. For example, if the standard deviation was 1 m, each point was shifted by a random error drawn from a normal distribution, with equal offsets applied to all three coordinate axes. Figure 9 presents the results of the noise sensitivity analysis, in which random error levels ranged from 0.2 m to 4 m. Most multiscale combinations maintained high accuracy, exceeding 90% when the noise level was around 1 m. In contrast, single-scale combinations using smaller voxel sizes showed higher sensitivity to noise. These results indicate that voxel-based aggregation helps suppress the influence of random errors on individual points, enhancing the reliability of supervoxel signatures under noisy conditions.

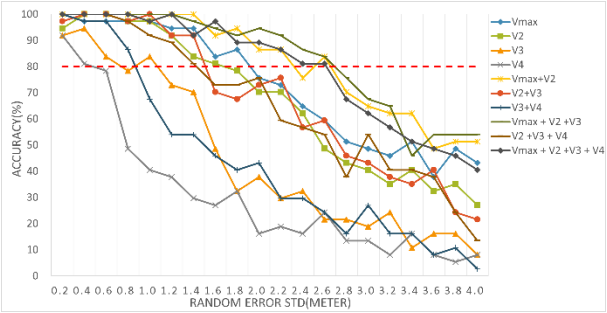

Figure 9. Matching accuracy of the noise test.

3.2 Real Case Study

We conducted a real cross-sensor matching experiment to validate the effectiveness of the proposed supervoxel signature method. The target point clouds were acquired using a Riegl VMX-250 mobile laser scanner for mapping purposes, while the sensed point clouds were obtained from a Velodyne's Puck lidar sensor mounted on a self-driving bus. These two datasets, representing cross-sensor data, partially overlapped in spatial coverage. We divided both datasets into 67 frames, each covering an area of 80 × 80 meters with a 25% overlap between adjacent frames. Notably, the cross-sensor point clouds were not subjected to any prior precise registration. As a result, the extracted patch pairs do not correspond exactly to the same spatial extent and are affected by slight rotational and translational discrepancies. Furthermore, the selected road segments were straight and homogeneous in structure, making the scenario particularly challenging due to the low geometric variability. Figure 10 shows an example frame pair, while Figure 11 illustrates part of the selected route segments. The data was collected in the Shuinan Trade and Economic Park, in Taichung City. The average point density for Riegl and Velodyne was 12986.63 pts/m^2 and 870.08 pts/m^2, respectively.

It is important to note that there were significant differences between these two datasets, not only in terms of point density but also in terms of random errors introduced by the sensors. We selected the same planar road surface to quantify the differences in random errors and calculated the standard deviation in the vertical direction. The target point cloud, which was scanned for mapping purposes (i.e., Riegl), exhibited a random error of only 0.04 m on the road surface. However, for the sensed point cloud (i.e., Velodyne) aimed at collecting real-time 3D information, the random error increased rapidly with an increase in distance from the sensor. This resulted in specific patterns in the point cloud, such as the ring pattern characteristic of the Velodyne sensor. The road surface of the sensed point cloud displayed a ring pattern, and the random error in that road surface was approximately 0.36 m due to the lack of a POS system.

In this real case study, we designed the matching evaluation procedure to reflect a practical application scenario. The centerlines of road segments were used as spatial references for voxelizing the point clouds, ensuring spatial continuity across adjacent scans. A trained DNN-AE was then applied to extract features and construct multiscale supervoxel signatures. For each segment, a search sequence was pre-constructed using supervoxel signature tracklets derived from the target point clouds. To evaluate the matching accuracy, each tracklet generated from the sensed point clouds was matched to the search sequence through a sliding window strategy.

Based on the simulation analysis, we selected the two most stable scale combinations— $V_{max} + V_2$ and $V_{max} + V_2 + V_3$ —for the sequence matching experiments. Different window sizes were defined to construct the supervoxel signature tracklets. When the window size was set to 1, the original per-frame supervoxel signature was used for matching. A window size of 3 indicates that a tracklet was formed by concatenating the current frame with its preceding and following frames. As shown in Table 3, the dual-scale combination Vmax+V2 consistently outperformed the multiscale combination Vmax+V2+V3. Specifically, when the window size was increased to 5, the matching accuracy exceeded 80%, and when extended to a window size of 9, the accuracy surpassed 90%. These results demonstrate that the proposed multiscale supervoxel signatures are capable of supporting reliable cross-sensor point cloud matching, even under challenging real-world conditions. However, it is important to note that supervoxel codes derived from smaller voxel scales may lead to over-segmentation of the point cloud, potentially amplifying noise and degrading feature stability.

Combination	Window Size	Accuracy(%)
$V_{max} + V_2$	1	49.25
	3	73.85
	5	82.54
	7	86.89
	9	91.53
	11	94.74
$V_{max} + V_2 + V_3$	1	35.82
	3	61.54
	5	71.43
	7	85.25
	9	91.53
	11	92.98

Table 3. Matching accuracy of the real-case datasets.

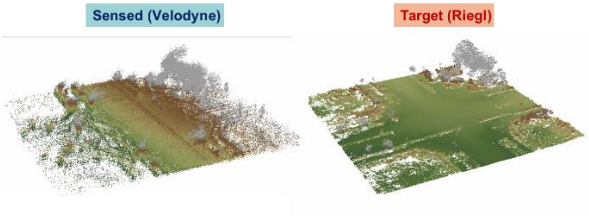

Figure 10. Real case datasets (S05&T05)

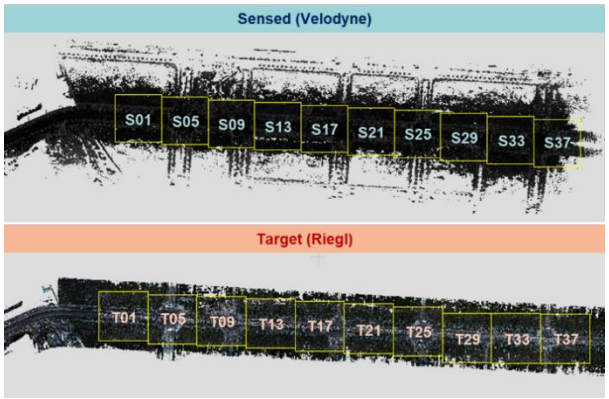

Figure 11. Real case datasets

3.3 Efficiency Discussion

The DNN-AE utilized in this study comprised two stages: the training and prediction stages. The training stage is time-consuming, making it challenging to achieve real-time processing. However, the prediction stage offers faster processing times compared to the training stage, allowing for near real-time implementation. Considering that the target point cloud was pre-scanned using Riegl HDLidar, the proposed matching method shows potential for initial alignment with VLP LDLidar. Furthermore, the data volume required for the supervoxel signatures was significantly smaller than that of the original lidar points. The compact representation of the supervoxel signature and the pre-trained encoder for HDLidar points only necessitated a minimal data volume for matching. This advantage underscores the effectiveness of utilizing supervoxel signatures for matching. This section highlights the proposed method's effectiveness for data compression using the real-case dataset. We used a personal computer with an i5-9600KF @3.70 GHz CPU for computation. Table 4 summarizes key metrics across all 10 frames, including the average number of points, point density, data capacity, and the data volume of multiscale supervoxel signatures.

This study aimed to significantly reduce the volume of 3D point cloud data while still achieving cross-sensor matching. Data compression was performed through a DNN-AE, and the compression ratio was the data volume between the supervoxel signature and the original point cloud (see Figure 12). Figure 13 shows one of the decompression results, which the DNN-AE predicted. The input features were supervoxels with voxel extraction features, which were compressed into supervoxel codes through the trained encoder. The advantage of using a DNN-AE is that the trained encoder can decompress the supervoxel codes back to the voxel's features. Therefore, supervoxel codes can be used for lidar matching and voxelized lidar data compression.

		Sensed	Target
Number of points		4.79×10^7	8.16×10^8
Capacity (GB)		1.76	26.30
Density (pts/m^2)		870.08	12986.63
Compressed volume (KB)	One Scale	0.41	
	Two Scales	0.81	
	Three Scales	1.22	
	Four Scales	1.62	

Table 4. Summary of point cloud statistics and compressed data volume for supervoxel signatures.

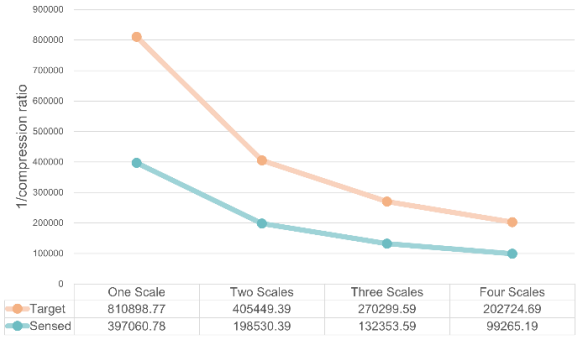

Figure 12. The compression ratio of multiscale supervoxel signatures.

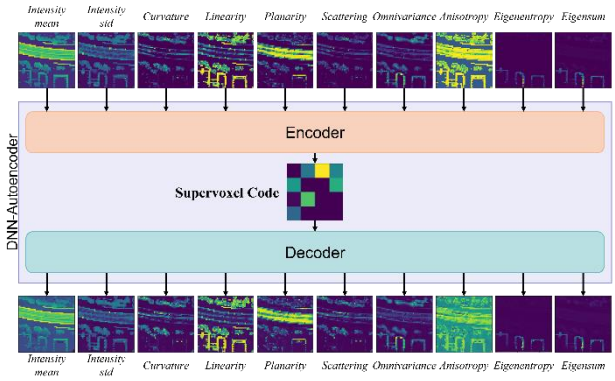

Figure 13. Illustration of compression and decompression by the proposed DNN-AE.

4. Conclusions and Future Work

This study presents an automatic framework for cross-sensor point cloud matching using a DNN-AE. The framework involves pre-processing the point clouds through voxelization and reprojection, extracting multiscale supervoxels from each point cloud, encoding the supervoxels using a DNN-AE, establishing supervoxel signatures by stacking multiscale supervoxel codes, and assessing the similarity of supervoxel signatures for cross-sensor lidar matching. The framework assumes that the point clouds acquired by different lidar systems have already been transformed into a common mapping system using POS information. The proposed supervoxel signature, a compressed representation set generated by a DNN-AE, exhibited tolerance for random errors and differences in point density between the lidar systems. The evaluation of the multiscale supervoxel signature demonstrates that combinations of dual-scale and multiscale configurations achieved higher matching accuracy under various simulation conditions. This finding was also supported by a real case study, where the supervoxel signature reduced the data volume while achieving an 80% matching accuracy. However, there are certain areas for improvement in the current framework. For a more comprehensive 3D description, voxelization along all three axes is necessary. Additionally, the proposed methods rely on specific conditions, such as similar intensity. Developing a more general framework would be a crucial direction for future work, which would include incorporating 3D voxelization and designing a method for testing the significance of each feature in supervoxels. Furthermore, an

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W6-2025-217-2025 | © Author(s) 2025. CC BY 4.0 License.

important goal for future research would be to enhance the design of supervoxel signatures to make it a more robust registration method, independent of initial conditions.

Acknowledgments

This research was partially supported by the Ministry of Interior, Taiwan, and National Science and Technology Council, Taiwan.

References

Besl, P.J., McKay, N.D., 1992: Method for registration of 3-D shapes. In: *Sensor Fusion IV: Control Paradigms and Data Structures*, Vol. 1611, 586–606. SPIE.

Elbaz, G., Avraham, T., Fischer, A., 2017: 3D point cloud registration for localization using a deep neural network auto-encoder. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 4631–4640.

Fu, K., Yuan, M., Liu, S., Wang, M., 2023: Boosting Point-BERT by multi-choice tokens. *IEEE Trans. Circuits Syst. Video Technol.*, 34(1), 438–447.

Huang, X., 2017: Learning a 3D descriptor for cross-source point cloud registration from synthetic data. *arXiv preprint* arXiv:1708.08997.

Huang, X., Mei, G., Zhang, J., Abbas, R., 2021: A comprehensive survey on point cloud registration. *arXiv preprint* arXiv:2103.02690.

Hui, L., Cheng, M., Xie, J., Yang, J., Cheng, M.M., 2022: Efficient 3D point cloud feature learning for large-scale place recognition. *IEEE Trans. Image Process.*, 31, 1258–1270.

Javanmardi, M., Javanmardi, E., Gu, Y., Kamijo, S., 2017: Towards high-definition 3D urban mapping: Road feature-based registration of mobile mapping systems and aerial imagery. *Remote Sens.*, 9(10), 975.

Jia, S., Liu, C., Wu, H., Huan, W., Wang, S., 2024: Incremental registration towards large-scale heterogeneous point clouds by hierarchical graph matching. *ISPRS J. Photogramm. Remote Sens.*, 213, 87–106.

Kim, G., Kim, A., 2018: Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map. In: *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 4802–4809.

Komorowski, J., 2021: MinkLoc3D: Point cloud based large-scale place recognition. In: *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 1790–1799.

Li, J., Zhan, J., Zhou, T., Bento, V.A., Wang, Q., 2022a: Point cloud registration and localization based on voxel plane features. *ISPRS J. Photogramm. Remote Sens.*, 188, 363–379.

Li, S., Ye, Y., Liu, J., Guo, L., 2022b: VPRNet: Virtual points registration network for partial-to-partial point cloud registration. *Remote Sens.*, 14(11), 2559.

Liu, Z., Zhou, S., Suo, C., Yin, P., Chen, W., Wang, H., ..., Liu, Y.H., 2019: LPD-Net: 3D point cloud learning for large-scale place recognition and environment analysis. *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2831–2840.

Milford, M.J., Wyeth, G.F., 2012: SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In: *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 1643–1649.

Nagy, B., Benedek, C., 2018: Real-time point cloud alignment for vehicle localization in a high resolution 3D map. In: *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 0–0.

Teo, T.A., Huang, S.H., 2014: Surface-based registration of airborne and terrestrial mobile LiDAR point clouds. *Remote Sens.*, 6(12), 12686–12707.

Uy, M.A., Lee, G.H., 2018: PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 4470–4479.

Wu, H., Scaioni, M., Li, H., Li, N., Lu, M., Liu, C., 2014: Feature-constrained registration of building point clouds acquired by terrestrial and airborne laser scanners. *J. Appl. Remote Sens.*, 8(1), 083587.

Xiong, B., Jiang, W., Li, D., Qi, M., 2021: Voxel grid-based fast registration of terrestrial point cloud. *Remote Sens.*, 13(10), 1905.

Xu, D., Liu, J., Hyyppä, J., Liang, Y., Tao, W., 2022: A heterogeneous 3D map-based place recognition solution using virtual LiDAR and a polar grid height coding image descriptor. *ISPRS J. Photogramm. Remote Sens.*, 183, 1–18.

Xu, M., Zhong, X., Zhong, R., 2025: A multi-source heterogeneous point cloud fine registration method for large-scale outdoor scenes. *IEEE Trans. Geosci. Remote Sens.*

Xu, Y., Tong, X., Stilla, U., 2021: Voxel-based representation of 3D point clouds: Methods, applications, and its potential use in the construction industry. *Autom. Constr.*, 126, 103675.

Yang, B., Zang, Y., Dong, Z., Huang, R., 2015: An automated method to register airborne and terrestrial laser scanning point clouds. *ISPRS J. Photogramm. Remote Sens.*, 109, 62–76.

Yang, J., Zhang, C.A., Wang, Z., Cao, X., Ouyang, X., Zhang, X., ..., Zhang, Y., 2024: 3D registration in 30 years: A survey. arXiv preprint arXiv:2412.13735.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J., 2022: Point-BERT: Pre-training 3D point cloud transformers with masked point modeling. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 19313–19322.

Zhang, D., Huang, T., Li, G., Jiang, M., 2012: Robust algorithm for registration of building point clouds using planar patches. *J. Surv. Eng.*, 138(1), 31–36.

Zhao, G., Guo, Z., Du, Z., Ma, H., 2025: Cross-PCR: A robust cross-source point cloud registration framework. *Proc. AAAI Conf. Artif. Intell.*, 39(10), 10403–10411.

Zou, X., Li, J., Wang, Y., Liang, F., Wu, W., Wang, H., ..., Dong, Z., 2023: PatchAugNet: Patch feature augmentation-based heterogeneous point cloud place recognition in large-scale street scenes. *ISPRS J. Photogramm. Remote Sens.*, 206, 273–292.