Cross-Sensor Robustness and Spatial Generalization for 3D Railway Point Cloud Semantic Segmentation

Arshia Ghasemlou[1], Mario Soilán[1], Jesús Balado[1], Belén Riveiro[1]

[1] CINTECX, Universidade de Vigo, GeoTECH Research Group, Campus Universitario de Vigo, As Lagoas, Marcosende, 36310 Vigo, Spain

Keywords: 3D Point-Clouds; Semantic Segmentation; Cross-sensor Generalization; Spatial Generalization; LiDAR; Railway Infrastructure

Abstract

Accurate semantic segmentation of 3D railway point clouds is essential for enabling automated inspection and asset management. Although recent deep learning (DL) models achieve strong performance on large benchmark datasets, their ability to generalize to point clouds captured with different sensors and in different spatial environments remains insufficiently explored. This study investigates the cross-sensor robustness and spatial generalization of state-of-the-art DL architectures for 3D semantic segmentation in railway scenarios. Three advanced models, Point Transformer V3, MinkUNet, and Swin3D, were trained on the SemanticRail3D dataset and evaluated on a newly acquired railway section scanned using three heterogeneous LiDAR systems: a terrestrial laser scanner (Faro Focus S150+), and two handheld mobile mapping devices (CHCNAV RS10 and GeoSLAM ZEB Go). The test area was manually annotated to provide high-quality ground truth for quantitative assessment.
Results show substantial performance variations across sensors, driven by differences in point density, noise levels, and scanning geometry. Domain-shift effects were evaluated directly from the model prediction outputs, including per-class IoU differences, uncertainty patterns, and cross-model agreement across sensors. To improve the robustness, an ensemble fusion strategy is evaluated to mitigate cross-sensor variability. The findings highlight the challenges of deploying DL models in real-world railway environments and provide insights for improving sensor-agnostic segmentation pipelines.

1. Introduction

Railway infrastructure plays a fundamental role in modern societies by facilitating both passenger mobility and freight transport, thereby supporting economic development and territorial cohesion. The safe and reliable operation of rail networks requires continuous inspection and maintenance activities, which represent a significant financial burden for infrastructure managers. In Europe alone, annual expenditure for inspection, monitoring, and maintenance of railway assets is estimated at €15–25 billion (Grandio et al., 2022). Given the extensive spatial scale of railway corridors and the high density of structural and operational elements, traditional inspection procedures—largely dependent on manual surveying—are increasingly unsustainable, inefficient, and difficult to scale to the demands of modern networks (Dekker et al., 2023).

Railway environments also present a unique set of perception challenges arising from their visually complex and highly repetitive structure. As highlighted by recent studies in automated inspection, even seemingly simple components such as clips, bolts, nuts, or fasteners become difficult to detect or classify due to background clutter, occlusions, material degradation, and strong illumination variability. These issues, while reported in 2D image-based inspection tasks, manifest even more severely in 3D point clouds where small or slender objects may disappear under sparse sampling, low incidence angles, or sensor-specific noise patterns. Moreover, the repetitive geometry of sleepers, ballast, rails, and catenary elements produces feature ambiguity, making it difficult for Deep Learning (DL) models to extract distinctive representations. Such factors indicate that railway scenes inherently amplify the sensitivity of DL models to sensor characteristics and acquisition conditions, thereby exacerbating the domain-shift problem addressed in this work (Mahamivanan et al., 2025).

2. Related Work

Recent advances in 3D sensing technologies, particularly LiDAR-based Mobile Mapping Systems (MMS) and UAV-mounted LiDAR, have transformed the digitalization of railway environments. These systems enable the rapid acquisition of high-resolution 3D point clouds that provide comprehensive geometric, structural, and contextual information on rails, catenary components, vegetation, signalling assets, and surrounding features (Grandio et al., 2022; Lamas et al., 2021). Such 3D representations underpin a wide range of applications, including asset inventory, condition assessment, clearance verification, predictive maintenance, and the creation of as-is Building Information Models (BIM).

The emergence of DL has further accelerated progress in this domain. Modern architectures, ranging from point-wise networks (PointNet, PointNet++), voxel-based models (SparseConv, MinkowskiNet), projection-based approaches (SalsaNext), to graph and transformer-based methods, have demonstrated strong performance in segmenting complex 3D scenes and outperform traditional heuristic or rule-based methods (Yarroudh et al., 2024). In railway environments, DL-based segmentation has shown high accuracy across linear and punctual elements such as rails, masts, wires, droppers, and signals, even in spatially extensive and cluttered scenarios (Grandio et al., 2022; Lamas et al., 2021).

However, despite these advances, a critical limitation persists: semantic segmentation models exhibit substantial performance degradation when applied to point clouds acquired with different sensors, acquisition geometries, or environmental conditions than those seen during training. This phenomenon, commonly referred to as domain shift, arises from variations in LiDAR characteristics such as beam configuration, scan pattern, intensity

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-XI-2-2026-85-2026 | © Author(s) 2026. CC BY 4.0 License.

behavior, viewpoint, trajectory, point density, and noise distribution (Rochan et al., 2022). Similar issues have been reported across autonomous driving datasets, where cross-sensor evaluation leads to significant accuracy loss even for state-of-the-art models.

Although domain shift has been acknowledged in other 3D perception fields, its specific manifestation in railway environments remains poorly understood. The combination of long linear geometries, narrow structures such as wires and droppers, highly repetitive elements, and platform-dependent occlusions makes railway point clouds particularly challenging for generalization. Furthermore, existing datasets often exhibit strong spatial autocorrelation: point distributions, catenary layouts, material properties, and environmental contexts are similar along long stretches of track. As a result, models may overfit to dataset-specific characteristics rather than learning sensor-invariant and scene-invariant representations. Understanding the mechanisms that cause generalization failures, and how they differ across sensors and environments, is therefore essential for deploying reliable segmentation models in operational railway monitoring (Soilán et al., 2019; Soum-Fontez et al., 2023; Wulff et al., 2024).

In the railway domain, this limitation is particularly pronounced. Railway point clouds may originate from a diverse set of platforms, including train-mounted MMS, trolley systems, backpack scanners, static terrestrial LiDAR, and UAV-borne sensors, each introducing distinct sampling patterns and sparsity levels. Moreover, railway corridors exhibit substantial spatial heterogeneity: geometric variability in track layouts, tunnel and bridge structures, catenary configurations, vegetation density, and regional construction standards all contribute to domain variability. As a result, models trained on a single sensor or geographic region frequently fail to generalize to new railway sections, limiting their operational applicability.

Existing public railway point-cloud datasets, including large-scale benchmarks, are predominantly acquired using a single sensor configuration, typically train-mounted mobile mapping systems. While these datasets are invaluable for training high-capacity models, they provide limited variability in terms of point density, incidence angles, scanning trajectories, and noise characteristics. Consequently, they do not adequately represent the diversity of data generated in real-world railway operations, where backpack scanners, handheld SLAM devices, UAV-LiDAR, and terrestrial static scanners are increasingly used. This lack of sensor diversity has created a critical gap in our understanding of how DL models behave when confronted with heterogeneous 3D data sources.

Despite growing interest in AI-driven railway analysis, very few studies systematically evaluate segmentation performance across multiple acquisition platforms or geographic regions. Most research focuses on algorithmic improvements or benchmarking within homogeneous datasets, without addressing how segmentation quality degrades under cross-sensor or cross-scene transfer. Moreover, existing evaluations typically consider only short track segments or controlled environments, limiting their applicability to the diverse conditions encountered in national rail networks. There is therefore a pressing need for a structured investigation that jointly analyses sensor variability, scene variability, and their combined impact on semantic segmentation robustness (Dekker et al., 2023; Ghasemlou et al., 2025a; Jiang et al., 2024; Kharroubi et al., 2024).

The importance of achieving cross-sensor robustness and spatial generalization is therefore twofold. First, for long-term adoption in Digital Twins, inventory management, and condition-monitoring workflows, segmentation models must remain reliable under heterogeneous sensing conditions without requiring exhaustive retraining or manual relabeling. Second, infrastructure operators increasingly rely on multi-source data integration, where consistent semantic labeling across sensors and time periods is essential for downstream analytics such as change detection and degradation modeling.

Despite the relevance of this challenge, systematic investigations of domain robustness in railway 3D segmentation remain scarce. Existing studies often rely on uniform datasets, focus on a limited set of sensors, or assess performance only on short sections of track with highly homogeneous characteristics (Grandio et al., 2022; Lamas et al., 2021; Rampriya et al., 2021). Although domain adaptation methods have recently been explored for autonomous driving LiDAR (Rochan et al., 2022), their application to railway environments, characterized by highly structured yet sensor-dependent spatial patterns, has not been extensively investigated.

In this context, the present work addresses the need for robust, sensor-invariant, and spatially generalizable point cloud semantic segmentation methods tailored to the complexities of railway environments. By focusing on cross-sensor and cross-region generalization, this study aims to bridge a critical gap between academic research and the operational deployment of large-scale, real-world 3D railway monitoring systems.

The contributions of this work are as follows:

1. **Systematic cross-sensor and cross-scene generalization study.**
We provide the first comprehensive evaluation of how state-of-the-art 3D point cloud semantic segmentation models generalize across heterogeneous railway environments. The study examines robustness under variations in sensing modality, acquisition geometry, spatial context, and point cloud quality, capturing the combined effects of both cross-sensor and cross-scene domain shift.

2. **Comparative analysis of three LiDAR acquisition systems.**
We assess the suitability of three distinct LiDAR technologies for semantic segmentation in railway scenarios:
(i) a terrestrial laser scanner Faro Focus S150+,
(ii) a handheld structured-light mobile mapping device CHCNAV RS10, and
(iii) a handheld SLAM-based LiDAR scanner GeoSLAM ZEB Go.
Their respective strengths, limitations, and interoperability are analysed in terms of density, noise characteristics, and spatial coverage.

3. **Creation of a manually annotated benchmark for domain-shift evaluation.**
A newly acquired 120-m railway scene is manually labelled to produce high-quality ground truth data, forming a benchmark that enables controlled and quantitative evaluation of cross-domain performance degradation. This dataset supports reproducibility and future research on domain generalization in railway mapping.

4. Inference-based analysis of domain shift.
Instead of relying on latent feature extraction, domain variability was analysed directly through the predicted outputs. Performance differences across devices were studied using per-class Intersection over Union (IoU), and cross-model agreement.

5. Ensemble fusion strategy for improving cross-domain robustness.

To alleviate accuracy drops caused by sensor- and scene-dependent differences, we implement an ensemble fusion method that integrates predictions from multiple DL models. This approach enhances segmentation stability and improves overall robustness under domain shift.

3. Dataset

3.1 SemanticRail3D Dataset

The models evaluated in this study were trained using the SemanticRail3D dataset, one of the largest publicly available benchmarks for semantic and instance segmentation of railway point clouds. The dataset was developed to address the scarcity of large, high-quality annotated point clouds in railway environments, a limitation that has hindered both the training of DL models and the establishment of common benchmarks for objective comparison. The dataset consists of 438 point clouds, each covering approximately 200 m of railway corridor, for a total of roughly 2.8 billion points. All data were captured using an Optech LYNX Mobile Mapper equipped with two LiDAR sensors, providing an average density of approximately 980 points/m² and a range precision of about 5 mm. Each point cloud is provided in a local, positive coordinate system, and includes several scalar attributes such as intensity , timestamp, return number, number of returns, and scan angle (–90° to 90°).

SemanticRail3D provides point-level annotations for 11 labelled classes, together with track-position information and instance segmentation. The semantic categories include: unclassified surfaces, rails, catenary components, contact wires, droppers, other overhead wires, masts, signs, traffic lights, track marks, signs mounted on masts, and lights. This detailed annotation scheme captures both the main structural elements of the railway infrastructure and finer operational components, enabling the development and evaluation of advanced semantic and instance segmentation methods. By offering a large, consistently labelled dataset acquired from a single high-precision MMS platform, SemanticRail3D provides a valuable benchmark while also highlighting the challenges of generalizing models beyond a uniform sensor configuration.

A key characteristic of SemanticRail3D is that it is captured entirely with a single sensor configuration, following a consistent acquisition trajectory. This ensures high internal uniformity, an advantage for training deep learning models, but also introduces strong spatial autocorrelation, as consecutive point clouds often share similar structural patterns, viewpoint geometry, and environmental context. While the scale and annotation quality make SemanticRail3D a valuable training resource, this uniformity may lead to dataset-specific biases, limiting the ability of models to generalize to point clouds acquired with different sensing modalities or in significantly different spatial environments (Ghasemlou et al., 2025; Ghasemlou et al., 2025b).

For these reasons, SemanticRail3D is particularly suitable for studying the domain-shift problem: it provides large, high-quality training data while simultaneously lacking sensor diversity and representation of alternative acquisition geometries, such as terrestrial laser scanning or handheld SLAM systems.. This characteristic makes it an ideal dataset for evaluating whether deep learning models trained on high-density MMS data can successfully transfer to lower-density, noisier, or structurally different point clouds, precisely the focus of this study.

3.2 Case Study Data

For this study, a railway section of approximately 120 meters in length was surveyed using three LiDAR sensors representing distinct acquisition technologies and point-cloud typologies. The objective of employing multiple sensors was to create a heterogeneous evaluation scenario that enables a rigorous assessment of the cross-sensor and cross-scene generalization capabilities of DL models. The selected devices include:
(1) a Faro Focus S150+ terrestrial laser scanner (TLS),
(2) a CHCNAV RS10 handheld mobile mapping system, and
(3) a GeoSLAM ZEB Go handheld SLAM-based LiDAR scanner.

These platforms differ substantially in terms of their acquisition principles, point density, noise characteristics, scanning geometry, and mobility. The Faro Focus S150+ provides high-accuracy static scans with extremely dense and low-noise point clouds, making it suitable as a reference for high-fidelity geometric documentation. In contrast, the CHCNAV RS10 captures data through a handheld structured-light trajectory, offering improved operability in narrow or obstructed spaces while maintaining moderate point density and relatively low noise levels. The GeoSLAM ZEB Go, based on simultaneous localization and mapping (SLAM), prioritizes fast, flexible data capture but typically produces sparser point clouds with higher noise due to motion-induced drift and trajectory estimation uncertainties.

Figure 1. LiDAR sensors employed in this work. (a) Faro S150+. (b) CHCNAV RS10. (c) Geoslam ZEB Go.

Figure 1 illustrates the three sensors used in this work, and Figure 2 displays the resulting colored 3D point clouds for the study site, highlighting the visual and structural differences across acquisition platforms. Table 1 provides a detailed comparison of the main technical specifications of the sensors, including range, accuracy, scan speed, field of view, total number of acquired points, and point density. A qualitative assessment shows clear differences among the datasets: the Faro Focus S150+ delivers the highest geometric fidelity and the most homogeneous density distribution; the CHCNAV RS10 produces a balanced compromise between mobility and quality; and the GeoSLAM ZEB Go results in lower density and higher noise, especially in regions with abrupt movements or limited loop closure.

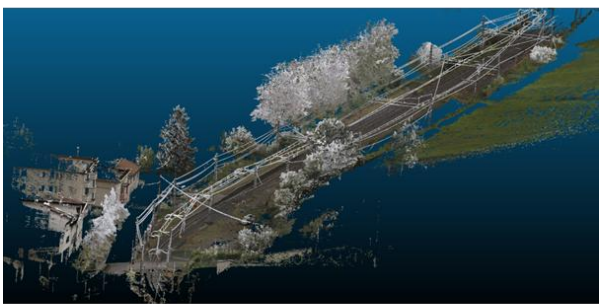
The quantitative values reported in Table 1 reinforce these observations. The TLS-based Faro dataset contains approximately 822.9 million points with a point density of 72,476

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-XI-2-2026-85-2026 | © Author(s) 2026. CC BY 4.0 License.

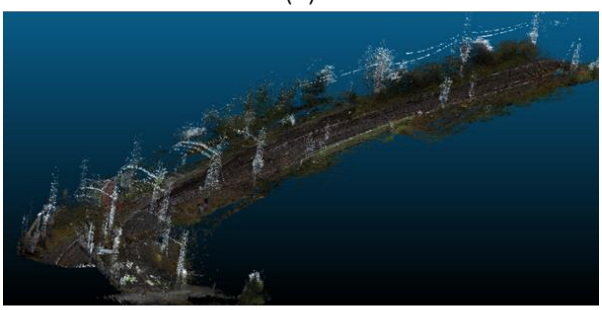
pts/m², enabling detailed modeling of fine railway elements. The CHCNAV RS10 acquisition contains 60.8 million points with a density of 11,826 pts/m², sufficient for reliable component segmentation. The GeoSLAM ZEB Go dataset is significantly sparser, with 4.5 million points and a density of 1,217 pts/m², representing a challenging scenario for DL models trained on higher-density data.

Together, these three datasets provide a comprehensive and diverse benchmark that captures the variability commonly encountered in real-world railway digitization workflows. This diversity allows us to rigorously evaluate the robustness and transferability of semantic segmentation models under heterogeneous sensing conditions.

(a)



(b)



(c)

Figure 2. 3D point clouds captured with (a) Faro S150+, (b) CHCNAV RS10, and (c) Geoslam ZEB Go.

3.3 Labelling of the Case Study Data

To generate reliable ground-truth annotations for the case study, the three point clouds (Faro Focus S150+, CHCNAV RS10, and GeoSLAM ZEB Go) were first registered into a common coordinate system. This alignment ensured that all datasets represented the same physical area despite differences in sensor geometry and acquisition coverage.

A shared 3D bounding box was then defined around the target 120-m railway segment and applied to all three aligned point clouds which is shown in Figure 3. This ensured consistent spatial extent and eliminated regions not captured by all sensors.

For semantic labeling, manual bounding boxes were drawn around each relevant railway object (e.g., rails, sleepers, masts, wires, signs). These boxes were propagated identically across the three point-clouds, and all points contained within each box were assigned the corresponding class label. Points outside all boxes were labelled as unclassified.

This process produced three consistent, object-level annotated datasets, one for each sensor, allowing fair comparison of segmentation performance under cross-sensor and cross-scene conditions.

Feature / Sensor	Faro Focus S150+	CHCNAV RS10	Geoslam Zeb GO
Type	TLS	Handheld	Handheld
Range	0.6 – 150m	0.05 – 120m	Up to 30m
Accuracy	±1 mm	<3 cm RTK	1–3 cm
Scan Speed	1M pts/s	1.92M pts/s	43,000 pts/s
Field of view	360º × 300º	360º × 270º	360° × 270°
Number of points	822.9M	60.8M	4.5M
Point density (pts/ m^2)	72,476	11,826	1,217

Table 1. Features and resolution metrics of the three LiDAR sensors employed in this work.

4. Methodology

4.1 Semantic Segmentation Algorithms

Semantic segmentation in this study was performed using three state-of-the-art architectures, Point Transformer v3, Swin3D, and MinkUNet, all implemented within the Pointcept framework (Pointcept Contributors, 2023). These architectures represent three complementary paradigms in 3D scene understanding: transformer-based attention networks, hierarchical windowed transformers, and sparse convolutional neural networks. Together, they provide a comprehensive basis for evaluating cross-sensor and cross-scene generalization in railway environments.

Point Transformer V3 employs a hierarchical self-attention mechanism combining positional encoding and multi-scale attention, enabling the model to capture both fine geometric structures and long-range dependencies. This makes it particularly suitable for segmenting elongated and thin railway elements such as catenary wires, contact lines, and droppers, where contextual relationships across different spatial scales are essential (Wu et al., 2024).

Swin3D extends the Swin Transformer to irregular 3D data using shifted window attention and a hierarchical representation. Its window-shifting strategy preserves feature continuity across boundaries while efficiently modelling both local and global context. This architecture excels in scenarios where repetitive patterns, such as sleepers, masts, or ballast surfaces, must be segmented consistently despite variations in density or viewpoint (Yang et al., 2025).

MinkUNet, based on the Minkowski Engine (an auto-differentiation library for sparse tensors), which leverages sparse 3D convolutions to process large-scale point clouds efficiently using voxelized representations. Its U-shaped encoder–decoder structure captures multi-scale features while maintaining computational efficiency, making it particularly well suited for dense railway environments. Unlike transformer-based models,

MinkUNet relies on convolutional locality, offering a complementary perspective in terms of robustness and sensitivity to sampling variations (Choy et al., 2019).

Across all models, lightweight data augmentation was applied during training, including random rotations, jittering, and point sampling variations, to improve robustness without disrupting the geometric integrity of railway structures. Training was conducted on a cluster equipped with five NVIDIA A100 GPUs, allowing for large batch sizes and stable convergence. Model performance was monitored on the SemanticRail3D validation set, and the best-performing checkpoints were selected for final evaluation. The final performance on the SemanticRail3D test set demonstrated strong segmentation accuracy across all three architectures, as shown in Table 2. These results confirm that all three architectures are highly competitive for high-resolution railway segmentation tasks while exhibiting distinct strengths depending on the nature of the input data.

Class	Swin 3D (IoU / Acc)	P.T. V3 (IoU / Acc)	MinkUNet (IoU / Acc)
Unclassified	**0.9977 / 0.9983**	0.9939 / 0.9968	0.9949 / 0.9980
Rail	**0.9372 / 0.9798**	0.8306 / 0.9082	0.8547 / 0.8997
Catenary	0.9125 / 0.9705	**0.9436 / 0.9802**	0.9378 / 0.9795
Contact	0.9249 / 0.9752	**0.9576 / 0.9869**	0.9552 / 0.9860
Droppers	**0.7337 / 0.8539**	0.7099 / 0.8126	0.6975 / 0.7960
Other wires	0.8389 / 0.9099	**0.8841 / 0.9309**	0.8669 / 0.9203
Masts	0.8902 / 0.9549	0.9122 / 0.9561	**0.9264 / 0.9642**
Signs	0.9231 / 0.9678	0.9191 / 0.9934	**0.9382 / 0.9631**
Traffic lights	0.7231 / 0.9729	**0.7711 / 0.9707**	0.7688 / 0.9525
Marks	0.7860 / 0.9096	0.8228 / 0.9484	**0.8456 / 0.9297**
Signs in masts	0.5338 / 0.6444	0.5101 / 0.7655	**0.5791 / 0.7538**
Lights	0.4351 / 0.4474	**0.7871 / 0.8071**	0.7761 / 0.8071
Overal	0.8030 / 0.9969	0.8368 / 0.9937	0.8451 / 0.9946

Table 2. Performance of the models on the Validation Dataset.

To assess cross-sensor and spatial generalization, the full preprocessing pipeline was applied to the three case-study datasets captured with the Faro Focus S150+, CHCNAV RS10, and GeoSLAM ZEB Go sensors. Inference across these heterogeneous point cloud typologies enables a detailed analysis of how each architecture responds to variations in point density, noise, acquisition geometry, and structural complexity. This evaluation provides deeper insight into the relative robustness of point-based, window-based, and sparse-convolutional models when deployed in real-world, multi-sensor railway inspection scenarios.

4.2 Inference on Case Study Data

To ensure compatibility between the case study point clouds and the SemanticRail3D trained models, a dedicated preprocessing workflow was applied to all three datasets. Since the acquisition characteristics, density, and spatial organization of the case study data differ significantly from those present in SemanticRail3D, these preprocessing steps were designed to normalize geometric patterns, reduce computational load, and maintain consistency with the model’s expected input distribution.
Normal estimation was performed for each point cloud using a k-nearest neighbours (k-NN) approach, providing local geometric context that facilitates the recognition of surfaces, edges, and fine structural components. Accurate normals are particularly important in railway scenes where many classes, such as rails, sleepers, or retaining walls, exhibit strong planar or linear geometric signatures.

To further homogenize the geometry, Principal Component Analysis (PCA) was applied to each point cloud. The principal direction was extracted and used as a reference axis to segment the railway corridor into six contiguous sections of approximately equal length. This segmentation strategy mirrors the linear organization of SemanticRail3D samples and effectively constrains the spatial extent processed in each forward pass, reducing GPU memory load and allowing the models to operate on structured, corridor-aligned subsets of the data.

Following segmentation, each subset was voxelized using a fixed voxel size of 2 cm, substantially reducing the number of points while preserving the structural detail necessary for semantic segmentation. This voxel resolution was selected to balance computational efficiency and geometric fidelity and matches the typical voxel scale used during the training stage.

For inference, a set of test-time augmentation (TTA) transformations was applied to improve model robustness and mitigate residual domain shift between the training data and the case study point clouds. Specifically, each segment was processed using five deterministic RandomScale factors (0.9, 0.95, 1.0, 1.05, and 1.1), generating scaled versions of the input while preserving the underlying geometry. In addition, the same five scaling factors were combined with a RandomFlip operation applied, producing mirrored variants of the point cloud. These augmentations were executed independently, resulting in ten forward passes per input segment. The final prediction was obtained by averaging the softmax probabilities across all augmentations and selecting the most probable class for each point.

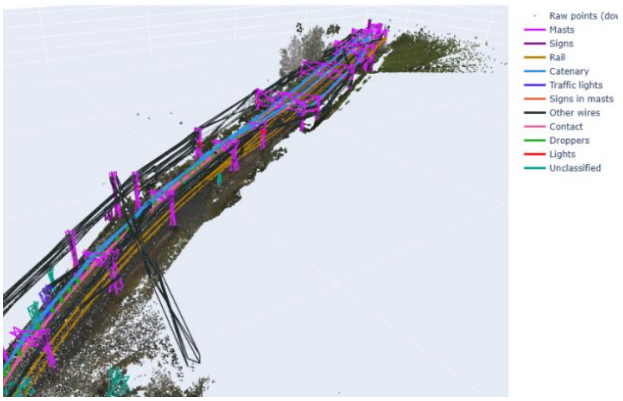


Figure 3. Bounding Boxes for Manual Labeling of the Case-Study Data.

This resulted in a total of 10 augmented forward passes per segment. For each prediction, class probabilities were obtained by applying a softmax layer, and the final per-point semantic label was computed by averaging the softmax outputs across all augmentations and selecting the class with the highest mean probability. This TTA strategy mitigates sensitivity to sensor-specific density variations, minor misalignments, and noise artifacts, factors that become especially relevant when transferring models trained on MMS data to handheld or SLAM-based point clouds.

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-XI-2-2026-85-2026 | © Author(s) 2026. CC BY 4.0 License.

The combination of preprocessing, PCA-based segmentation, voxelization, and augmented inference provides a standardized and robust pipeline for evaluating how the three models, Point Transformer V3, Swin3D, and MinkUNet, respond to geometric variability and acquisition heterogeneity in real-world railway environments.

4.3 Point Cloud Characteristics Analysis

Before presenting the inference results on the case-study datasets, it is important to quantify how their geometric and semantic characteristics differ from those of the SemanticRail3D training data. Although the same preprocessing pipeline was applied to all datasets, the three acquisition systems produce point clouds with markedly different sampling patterns, spatial extents, and scene compositions. Table 3 summarizes the main characteristics of the training split and the case-study datasets in terms of point count, density, vertical extent, and semantic diversity. The results reveal a clear acquisition-domain shift: both CHCNAV and Faro exhibit substantially higher point densities than the training data, while ZEB Go contains fewer points overall and a more compact spatial structure. Differences are also evident in semantic composition, with CHCNAV showing the highest class entropy, indicating a richer and more heterogeneous scene content, whereas ZEB Go exhibits the lowest semantic diversity.

Dataset	pts/m²	pts/m³	Class Entropy
SemanticRail3D	167	7.12	0.281
CHCNAV	2091	161.34	0.383
Faro	1665	85.15	0.281
ZEB Go	842	104.34	0.129

Table 3. Summary of point cloud characteristics across datasets.

These differences are further illustrated in Figure 4, which shows the distribution of the median first nearest-neighbour (1-NN) distance across datasets as a descriptor of local point spacing. Lower 1-NN values correspond to denser local sampling, whereas higher values indicate more widely spaced neighbourhoods. Compared with the training data, the case-study point clouds generally present denser local geometric structures, although the three devices retain distinct acquisition signatures. Such variations are expected to influence both inference efficiency and segmentation behaviour, since the evaluated architectures rely on local geometric relationships in different ways. Consequently, the characterization presented in Table 3 and Figure 4 provides an important context for interpreting the runtime, generalization, and ensemble results discussed in the following sections.

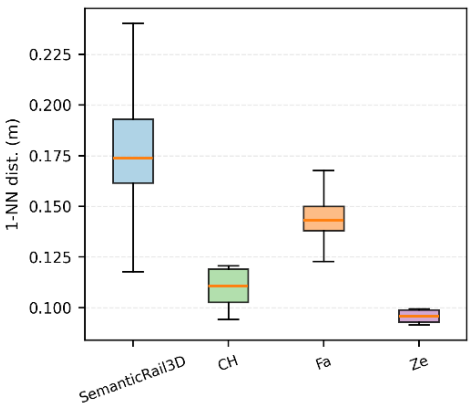


Figure 4. Local density comparison across datasets.

4.4 Inference computational cost analysis

To quantify the computational cost of the three architectures, we clocked the inference time on the test segments for each device. Table 4 summarises these results using four metrics: (i) the total number of points processed across the six segments for each device, (ii) the total clock inference time, obtained by summing the segment-level times, (iii) the average inference time per segment, and (iv) a normalized runtime, expressed as seconds per million points (s/Mpts), computed by dividing the total clock time by the total number of processed points (in millions). This normalization enables a fair comparison across sensors. As expected, denser Faro scans lead to longer runtimes, while the s/Mpts metric reveals inherent architectural differences: Point Transformer V3 exhibits the highest processing efficiency, followed by MinkUNet, whereas Swin3D consistently requires more time per million points. Overall, these results reflect how model design, scene complexity, and sensor-specific characteristics jointly influence inference speed.

Model	Device	# Points [M]	Inf. time [s]	Avg Inf. time / seg [s]	s/Mpts
Swin3D	CHCNAV	8.46	1187.7	197.95	140.31
	ZebGo	2.43	484.68	80.78	199.61
	Faro	9.33	2698.25	449.71	289.21
MinkUNet	CHCNAV	8.46	658.83	109.81	77.83
	ZebGo	2.43	277.45	46.24	114.26
	Faro	9.33	927.98	154.66	99.46
PTv3	CHCNAV	8.46	587.39	97.9	69.39
	ZebGo	2.43	199.17	33.19	82.02
	Faro	9.33	790.58	131.76	84.74

Table 4. Inference computational cost analysis over the three models across the case-study datasets.

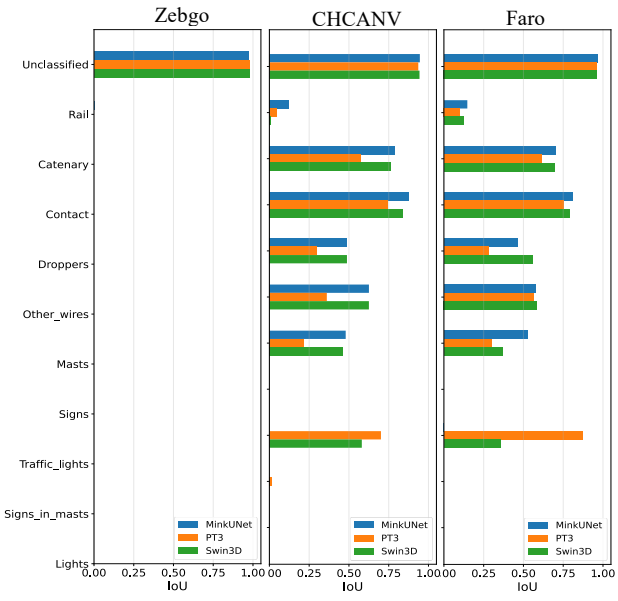


Figure 5. Generalization Results on the Case Study Data.

4.5 Generalization Results on the Case Study Data

Figure 5 shows the per-class IoU obtained by the three models when applied to the Faro case-study dataset. The plot summarizes how well each architecture generalizes from the training domain to this new sensor and scene. Several classes achieve valid IoU values, while others appear with zero IoU because they are not present in the case-study area and therefore cannot be evaluated.

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-XI-2-2026-85-2026 | © Author(s) 2026. CC BY 4.0 License.

Overall, the figure provides a concise overview of model behavior across the available classes and highlights the variability in generalization performance across different semantic categories.

4.6 Class-wise Expert Ensemble

To exploit the complementary strengths of the three architectures, a class-wise reliability-weighted ensemble was designed. Instead of treating all models equally, the proposed fusion strategy assigns a semantic-class-dependent reliability weight to each model based on its validation performance on the SemanticRail3D dataset.

Let $M = \{1,2,3\}$ denote the set of models corresponding to Swin3D, Point Transformer V3, and MinkUNet, and let Cbe the set of semantic classes. For each model $m \in M$ and class $c \in C$, a reliability weight $w_{m,c}$ is defined as the class-wise IoU obtained on the SemanticRail3D validation set:

$$w_{m,c} = \mathrm{IoU}_{m,c} \tag{1}$$

At inference time, each model predicts one semantic label for every point p, denoted as $\hat{y}_m(p) \in C$. For each candidate class c predicted by at least one model, a point-wise ensemble score is computed as

$$S(p,c) = \sum_{m \in M} w_{m,c}\ \mathbf{1}[\hat{y}_m(p) = c] \tag{2}$$

where $1[\cdot]$ is the indicator function, equal to 1 if model $m$predicts class c for point p, and 0 otherwise.

The final ensemble label for point p is then selected as the class with the maximum weighted score:

$$\hat{y}_{\mathrm{ens}}(p) = \arg\max_{c \in C} S(p,c) \tag{3}$$

This formulation can be interpreted as a class-wise weighted voting scheme, where votes from models that are more reliable for a given semantic category contribute more strongly to the final decision. In contrast to standard majority voting, the proposed approach accounts for the fact that different architectures exhibit different strengths across classes. For example, one model may be more reliable for linear overhead components such as contact wires, whereas another may perform better on volumetric or sparse objects such as masts or signs.

The proposed ensemble is computationally simple, requiring only the per-point predicted labels and the fixed class-wise IoU table derived from validation data. At the same time, it provides a principled way to transfer model-specific strengths into a single fused prediction for heterogeneous railway point clouds acquired with different sensors. The mIoU results of the individual models and the proposed ensemble are summarized in Table 5.

Device	Swin3D	PTv3	Mink	Ensemble
Chcnav	0.392	0.365	0.393	0.425
ZebGo	0.122	0.122	0.122	0.122
Faro	0.37	0.419	0.381	0.403

Table 5. Comparison of mIoU for each model and the proposed IoU-weighted ensemble across the three case-study datasets.

5. Discussion

The results indicate that cross-sensor transfer is influenced not only by domain shift, but also by the inherent differences in point cloud specifications between the training and inference datasets. Although the same preprocessing pipeline was applied to all case-study data, the three acquisition systems still produce point clouds with substantially different density, local sampling structure, noise characteristics, and semantic observability. This is also supported by the density and 1-NN analyses, which showed clear differences between the SemanticRail3D training data and the three case-study datasets. Such variations affect how local neighbourhoods and geometric patterns are represented, and therefore provide an additional explanation for why the achieved mIoU on the case-study datasets remains considerably lower than on the validation data.

A clear sensor-dependent pattern is also observed across the three case-study datasets. In the CHCNAV and Faro datasets, all three models were able to recover several important railway classes with meaningful consistency. In contrast, the GeoSLAM ZEB Go dataset showed much more limited generalization. Inspection of the confusion matrices and the manually labelled subsets indicates that several thin and overhead classes were either absent, severely underrepresented, or poorly captured in the ZEB Go data. As a result, the strong performance degradation on this dataset reflects not only model limitations, but also sensing and observability constraints.

A class-dependent trend is also evident. Larger and more geometrically stable classes, such as Unclassified and Masts, were generally transferred more reliably across devices, whereas smaller and thinner objects remained much more sensitive to acquisition quality and local geometric fidelity. This is especially visible for categories such as Droppers, Signs in masts, Lights, and in some cases Rail, which showed unstable behaviour across models and sensors. These findings suggest that semantic generalization in railway environments depends strongly on whether a class remains sufficiently represented after downsampling, voxelization, and sensor-specific degradation.

The three architectures also exhibited complementary strengths rather than a uniform ranking. The proposed IoU-weighted ensemble further showed that combining these complementary predictions can improve robustness in some cases, although it cannot fully overcome severe acquisition mismatch. Overall, the results demonstrate that evaluating railway point cloud segmentation under real deployment conditions requires not only model comparison, but also explicit consideration of sensor visibility, acquisition-driven bias, and the mismatch between training and inference point cloud characteristics.

6. Conclusion

This paper investigated the cross-sensor robustness and spatial generalization of state-of-the-art DL models for semantic segmentation of 3D railway point clouds. Three architectures, Point Transformer V3, Swin3D, and MinkUNet were trained on the large-scale SemanticRail3D benchmark and evaluated on a newly acquired 120 m railway section captured by three heterogeneous LiDAR systems: a terrestrial laser scanner (Faro Focus S150+), a handheld structured light device (CHCNAV RS10), and a handheld SLAM-based sensor (GeoSLAM ZEB Go). The case-study data were carefully registered, normalized, voxelized, and manually annotated to provide a consistent multi-sensor benchmark for quantitative evaluation.

The results confirm that all three models achieve strong performance on the training domain but experience noticeable performance degradation when transferred to new sensors and spatial contexts. The magnitude of this degradation varies across devices and semantic classes, reflecting differences in point density, noise level, and acquisition geometry. High-fidelity TLS data generally support better generalization than sparser, noisier

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-XI-2-2026-85-2026 | © Author(s) 2026. CC BY 4.0 License.

SLAM acquisitions, while thin or small objects such as wires and droppers remain particularly sensitive to changes in sampling characteristics. The analysis of per-class IoU and cross-model behaviour shows that no single architecture is uniformly superior; instead, the models exhibit complementary strengths across different object types.

An IoU-weighted ensemble strategy, informed by validation performance on SemanticRail3D, was proposed to exploit this complementarity. The ensemble consistently matches or improves the mIoU of the individual models on the case-study datasets, especially for the medium- and high-quality sensors, demonstrating that simple fusion schemes can enhance robustness under domain shift without additional training. The computational cost analysis further highlights trade-offs between accuracy and efficiency, with transformer-based models offering strong performance at a higher inference cost, and sparse-convolutional networks providing competitive accuracy with improved runtime characteristics.

Overall, the study underscores that cross-sensor and spatial generalization remain open challenges for deploying semantic segmentation models in operational railway monitoring. Future work will focus on extending the analysis to additional railway sites and sensor types, integrating explicit domain adaptation or domain generalization techniques, and incorporating uncertainty-aware outputs at object level to better support decision-making in digital twin and condition-monitoring applications.

Acknowledgements

This work has been supported by the Spanish Ministry of Science, Innovation and Universities through grant RYC2021-033560-I funded by MCIN/AEI/10.13039/501100011033; and by European Union NextGenerationEU/PRTR; and by grant ED431F 2024/02 funded by Xunta de Galicia, Spain-GAIN; and by RYC2022-038100-I funded by MICIU/AEI/10.13039/501100011033 and by ESF +. Additionally, some computational resources were provided by the Centro de Supercomputación de Galicia (CESGA).

References

Choy, C., Gwak, J., & Savarese, S. (2019, June). 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dekker, B., Ton, B., Meijer, J., Bouali, N., Linssen, J., & Ahmed, F. (2023). Point Cloud Analysis of Railway Infrastructure: A Systematic Literature Review. *IEEE Access*, *11*, 134355–134373. https://doi.org/10.1109/ACCESS.2023.3337049

Ghasemlou, A., Soilán, M., Martínez-Sánchez, J., Arias, P., Lorenzo, H., & Riveiro, B. (2025). *SemanticRail3D: A 3D Point Cloud dataset with semantic annotations of railway environments*. https://doi.org/10.5281/zenodo.15641832

Ghasemlou, A., Soilán, M., & Riveiro, B. (2025a). Semantic segmentation of imbalanced 3D point clouds in railway environments: comparative analysis of algorithms and training pipelines for semantic segmentation. *EG-ICE 2025*. https://doi.org/10.17868/strath.00093262

Ghasemlou, A., Soilán, M., & Riveiro, B. (2025b). SemanticRail3D - A Mobile LiDAR Benchmark for Semantic and Instance Segmentation of Railway Corridors. *Scientific Data 2025 13:1*, *13*(1), 82-. https://doi.org/10.1038/s41597-025-06392-9

Grandio, J., Riveiro, B., Soilán, M., & Arias, P. (2022). Point cloud semantic segmentation of complex railway environments using deep learning. *Automation in Construction*, *141*, 104425. https://doi.org/10.1016/J.AUTCON.2022.104425

Jiang, T., Li, S., Zhang, Q., Wang, G., Zhang, Z., Zeng, F., An, P., Jin, X., Liu, S., & Wang, Y. (2024). RailPC: A large-scale railway point cloud semantic segmentation dataset. *CAAI Transactions on Intelligence Technology*, *9*(6), 1548–1560. https://doi.org/10.1049/cit2.12349

Kharroubi, A., Ballouch, Z., Hajji, R., Yarroudh, A., & Billen, R. (2024). Multi-Context Point Cloud Dataset and Machine Learning for Railway Semantic Segmentation. *Infrastructures*, *9*(4), 71. https://doi.org/10.3390/infrastructures9040071

Lamas, D., Soilán, M., Grandío, J., & Riveiro, B. (2021). Automatic Point Cloud Semantic Segmentation of Complex Railway Environments. *Remote Sensing*, *13*(12), 2332. https://doi.org/10.3390/rs13122332

Mahamivanan, H., Matthews, J., Love, P. E. D., & Nasirzadeh, F. (2025). Toward accurate detection of small objects in rail construction: A deep learning perspective. *Engineering Applications of Artificial Intelligence*, *160*, 111977. https://doi.org/10.1016/j.engappai.2025.111977

Pointcept Contributors. (2023). *Pointcept: A Codebase for Point Cloud Perception Research*.

Rampriya, R. S., Sabarinathan, & Suganya, R. (2021). RSNet: Rail semantic segmentation network for extracting aerial railroad images. *Journal of Intelligent & Fuzzy Systems*, *41*(2), 4051–4068. https://doi.org/10.3233/JIFS-210349

Rochan, M., Aich, S., Corral-Soto, E. R., Nabatchian, A., & Liu, B. (2022). Unsupervised Domain Adaptation in LiDAR Semantic Segmentation with Self-Supervision and Gated Adapters. *2022 International Conference on Robotics and Automation (ICRA)*, 2649–2655. https://doi.org/10.1109/ICRA46639.2022.9811654

Soilán, M., Sánchez-Rodríguez, A., del Río-Barral, P., Perez-Collazo, C., Arias, P., & Riveiro, B. (2019). Review of Laser Scanning Technologies and Their Applications for Road and Railway Infrastructure Monitoring. *Infrastructures*, *4*(4), 58. https://doi.org/10.3390/infrastructures4040058

Soum-Fontez, L., Deschaud, J.-E., & Goulette, F. (2023). MDT3D: Multi-Dataset Training for LiDAR 3D Object Detection Generalization. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5765–5772. https://doi.org/10.1109/IROS55552.2023.10341614

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., & Zhao, H. (2024). *Point Transformer V3: Simpler, Faster, Stronger*. https://arxiv.org/abs/2312.10035

Wulff, F., Schäufele, B., Pfeifer, J., & Radusch, I. (2024). Railway LiDAR semantic segmentation based on intelligent semi-automated data annotation. *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*, 1–7. https://doi.org/10.1109/VTC2024-Fall63153.2024.10758029

Yang, Y. Q., Guo, Y. X., Xiong, J. Y., Liu, Y., Pan, H., Wang, P. S., Tong, X., & Guo, B. (2025). Swin3D: A pretrained transformer backbone for 3D indoor scene understanding. *Computational Visual Media*, *11*(1), 83–101. https://doi.org/10.26599/CVM.2025.9450383

Yarroudh, A., Kharroubi, A., Jeddoub, I., Ballouch, Z., & Billen, R. (2024). Railway reconstruction from 3D point cloud using Deep Learning and Parametric Modeling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLVIII-2/W8-2024*, 477–482. https://doi.org/10.5194/isprs-archives-XLVIII-2-W8-2024-477-2024