HOKKAIDO UNIVERSITY

Title	Transfer Learning for Infrastructure-Oriented Semantic Segmentation of Urban 3D Point Clouds
Author(s)	Ren, Kewei; Itakura, Kenta; Chun, Pang-jo et al.
Relation	Proceedings of the 10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026), June 28 - July 1, 2026, Sapporo, Japan
Citation	Proceedings of the 10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026)
Issue Date	2026-07-01
Doc URL	https://hdl.handle.net/2115/100467
Type	conference paper
File Information	C000190_1.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026),
28 June-01 July 2026, Sapporo, Japan
Kitahara, T., Miyamori, Y. & Beer, M. (editors)

Transfer Learning for Infrastructure-Oriented Semantic Segmentation of Urban 3D Point Clouds

Kewei Ren[1], Itakura Kenta[2], and Pang-jo Chun[3,] *

[1]*Department of Civil Engineering, Graduate School of Engineering, The University of Tokyo, Tokyo,113-0013, Japan*
E-mail: renkw@g.ecc.u-tokyo.ac.jp
[2]*ImVisionLabs Inc., Tokyo,113-0023, Japan.*
E-mail: kentai@imvisionlabs.com
[3]*Department of Civil Engineering, Graduate School of Engineering, The University of Tokyo, Tokyo, 113-0013, Japan.*
E-mail: chun@i-con.t.u-tokyo.ac.jp
**Corresponding Author*

Accurate semantic segmentation of urban 3D point clouds is a prerequisite for infrastructure condition monitoring and maintenance risk assessment. However, infrastructure-orientated urban scenes present segmentation challenges due to unclear boundaries with geometrically similar classes, high scene variability, and class imbalance in training datasets. Furthermore, training deep learning models such as the Point Transformer framework from scratch requires substantial computational resources and large-scale annotated datasets, which are often unavailable in real-world applications. To address these issues, this study proposes a transfer learning approach that uses pre-trained weights from the nuScenes autonomous driving dataset. Since the nuScenes dataset contains extensive urban driving scenes, the pre-trained model learns general geometric and contextual features from complex urban environments. Fine-tuning these models on target datasets reduces training time and computational costs while improving segmentation accuracy, which makes deep learning techniques more accessible for civil engineering practitioners. The proposed method builds upon the Point Transformer V3 architecture and integrates multi-scale context aggregation with boundary-aware learning to improve segmentation performance in ambiguous regions. Experimental results on a urban 3D point clouds segmentation task demonstrate the potential of transfer learning to bridge the gap between autonomous driving and civil infrastructure applications, achieving higher segmentation accuracy than training from scratch with considerably less training time, particularly for difficult and underrepresented classes such as outdoor ground and indoor ground surface.

Keywords: Point Cloud, Semantic Segmentation, Infrastructure condition monitoring, Road Surfaces.

1 Introduction

Three-dimensional point cloud data acquired by Light Detection and Ranging (LiDAR) sensors are increasingly used for infrastructure inspection, digital-twin development, and maintenance-oriented asset management. Semantic segmentation of point clouds—assigning a class label to every point—enables automated scene interpretation and provides the basis for downstream engineering tasks. In civil infrastructure applications, recent studies have shown that point-cloud semantic segmentation can support practical tasks such as bridge component recognition, geometric analysis, and automated dimension estimation, demonstrating its value beyond generic scene understanding (Chen et al. 2025; Lin et al. 2025).

Deep learning for point cloud processing has progressed from point-based methods such as PointNet (Qi et al. 2017a) and PointNet++ (Qi et al. 2017b), through convolution-based architectures including KPConv (Thomas et al. 2019), MinkowskiNet (Choy et al. 2019), and DGCNN (Wang et al. 2019), to transformer-based approaches such as Point Transformer (Zhao et al. 2021), Point Transformer V2 (Wu et al. 2022), and Point Transformer V3 (Wu et al. 2024). Among them, Point Transformer V3 (PTv3) achieves strong performance and scalability through a simplified serialized design, making it a promising backbone for large-scale semantic segmentation tasks (Wu et al. 2024). However, these methods remain general-purpose architectures and are not explicitly tailored to the ambiguities and class distributions commonly encountered in infrastructure-oriented urban scenes. Although image–point cloud fusion can enhance semantic understanding by complementing geometric information with dense visual cues, its effectiveness depends on accurate cross-modal alignment (Lin et al. 2025). In practical infrastructure scenarios, however, acquiring and maintaining well-aligned image and point cloud data can be challenging. This makes it still important to investigate geometry-based methods that rely solely on point cloud information.

In the present study, the reliable distinction between Outdoor Ground and Indoor Ground Surface is particularly important for infrastructure-oriented urban scene understanding in the context of civil engineering. Although these two classes often exhibit similar local geometric characteristics, they differ in their broader spatial contexts, environmental settings, and engineering interpretations. Misclassification between them can therefore reduce the reliability of semantic scene parsing in mixed urban environments and diminish the practical value of segmented point clouds for subsequent infrastructure analysis and management tasks. This challenge is especially pronounced in the Saitama urban

point-cloud dataset, which was previously used for transformer-based classification of urban infrastructure scenes and includes classes such as ground surfaces, buildings, utility poles and wires, vegetation, and vehicles (Itakura et al. 2025). In the present task setting, Outdoor Ground and Indoor Ground Surface form a particularly difficult class pair given the highly imbalanced class distribution. Accordingly, enhancing the separation of these two ground-related classes represents a central objective of this study.

Transfer learning provides a practical solution under this problem setting. Large-scale autonomous-driving datasets such as nuScenes (Caesar et al. 2020) and SemanticKITTI (Behley et al. 2019) have substantially advanced 3D scene understanding by providing diverse urban LiDAR data and large annotated corpora. Although their label taxonomies do not directly match infrastructure-oriented monitoring tasks, they still encode transferable geometric and contextual priors from complex urban scenes. Such pre-trained representations are especially valuable when the target dataset is relatively limited and when training a large transformer backbone entirely from scratch would require substantial computational resources (Caesar et al. 2020; Fang et al. 2024).

However, the present task remains challenging even with a strong backbone. First, local geometric cues alone are often insufficient to distinguish Outdoor Ground from Indoor Ground Surface when both classes exhibit similar planar structures. Second, boundaries between these classes can be geometrically smooth or visually subtle, reducing boundary precision in standard segmentation pipelines. Third, feature overlap between confusing classes, together with class imbalance, can hinder discriminability and stable optimization. These challenges are consistent with the observations in the later experimental sections of this study, where boundary ambiguity, confusing class pairs, and the underrepresented Indoor Ground Surface category emerge as key sources of error.

To address these challenges, this study proposes PTv3-Enhanced-V2, which initializes a PTv3 backbone with nuScenes pre-trained weights and augments it with three task-specific modules: (1) a Multi-Scale Context Aggregation (MSCA) module with channel attention and gated fusion to strengthen contextual reasoning across multiple spatial scales; (2) a Supervised Contrastive Learning (SCL) module with EMA-updated class prototypes and hard negative mining to improve feature-space separability; and (3) a Boundary-aware (BA) module that emphasizes ambiguous transition regions during training and inference. A composite loss combining cross-entropy, Lovász-Softmax (Berman et al. 2018), boundary-weighted supervision, contrastive learning (Khosla et al. 2020), and auxiliary boundary detection is adopted to jointly improve segmentation accuracy, boundary precision, and feature discriminability.

2 Related Work

Early point cloud segmentation methods operated directly on unordered point sets. PointNet (Qi et al. 2017a) achieved permutation invariance via shared MLPs and symmetric pooling; PointNet++ (Qi et al. 2017b) added hierarchical set abstraction for multi-scale local geometry. Convolution-based methods—KPConv (Thomas et al. 2019), MinkowskiNet (Choy et al. 2019), and and DGCNN (Wang et al. 2019)—improved feature extraction through kernel-based, sparse, and graph-convolutional designs.)—improved feature extraction through kernel-based, sparse, and random-sampling designs. Transformer architectures model long-range dependencies through self-attention: Point Transformer (Zhao et al. 2021) introduced vector self-attention for point clouds, PTv2 (Wu et al. 2022) improved efficiency through grouped attention, and PTv3 (Wu et al. 2024) achieved a three-fold speedup via serialized attention along space-filling curves. Despite this progress, existing architecture remain general-purpose and do not address boundary ambiguity or class imbalance in infrasture-oriented urban point cloud segmentations.

Transfer learning reduces data and computational requirements by reusing representations from large-scale source domains (Fang et al. 2024). Autonomous driving datasets such as nuScenes (Caesar et al. 2020) and SemanticKITTI (Behley et al. 2019) provide rich urban scene annotations but label road surfaces as a single class, lacking the granularity needed for infrastructure monitoring. The present work adapts nuScenes pre-trained representations to fine-grained surface labels through task-specific modules.

Loss-based and boundary-aware enhancements are also relevant. The Lovász-Softmax loss (Berman et al. 2018) directly optimizes the IoU metric, and the supervised contrastive loss (Khosla et al. 2020) separates class embeddings in feature space. Boundary-aware strategies, including auxiliary detection heads and spatially varying loss weights, direct model capacity toward ambiguous transitions. Our framework integrates these complementary strategies into a unified architecture for road surface segmentation.

3 Methodology

3.1 *Overall Framework and Transfer Learning Strategy*

The PTv3-Enhanced-V2 framework has three stages: (1) a PTv3 multi-scale backbone pre-trained on nuScenes (Caesar et al. 2020) extracts point-level features; (2) three enhancement modules—MSCA, SCL, and BA—process the decoder output; (3) a composite loss jointly optimizes segmentation accuracy, feature discriminability, and boundary precision.

The backbone is initialized with nuScenes weights encoding road geometry, surface continuity, and spatial relationships among urban objects. A differential learning rate is adopted: 5×10^{-5} for the backbone and 5×10^{-4} for the new modules, ensuring pre-trained features are refined gradually. Training runs for 100 epochs with AdamW (weight decay 0.05, gradient clipping 1.0) and a OneCycleLR schedule with 5% warmup and cosine annealing.

3.2 *PTv3 Multi-Scale Backbone*

The backbone adopts the PTv3 architecture (Wu et al. 2024) with multi-scale serialized attention. Each input point cloud carries six channels (XYZ + RGB) and is serialized along four space-filling curve orders (z-order, transposed z-order, Hilbert, transposed Hilbert). The encoder has five stages with channel dimensions (32, 64, 128, 256, 512) and depths (2, 2, 2, 6, 2); the decoder has four stages with dimensions (64, 64, 128, 256). Multi-scale attention computes self-attention at two patch sizes simultaneously within each stage, and outputs are fused by element-wise addition.

3.3 *Multi-Scale Context Aggregation*

Different road surface types share similar local geometry and require broader contextual cues for disambiguation. The MSCA module aggregates features at three spatial scales. Given decoder output $F \in \mathbb{R}^{N\times C}$, the module routes F through three parallel branches, applies channel attention, and merges outputs via gated fusion with learnable softmax weights:

$$F_{\text{out}} = F + \text{Gate}\left(F_{s_1}, F_{s_2}, F_{s_3}\right) \quad (1)$$

3.4 Supervised Contrastive Learning

The Supervised Contrastive Learning (SCL) module projects decoder features into a 64-dimensional L2-normalized embedding space. For each class, an EMA-updated prototype (momentum = 0.999) serves as a stable anchor. The contrastive loss encourages each embedding z to be similar to its class prototype p+ and dissimilar to others, with temperature $\tau = 0.5$. Hard negative mining (ratio 0.3) focuses on confusing class pairs. A 20-epoch warmup delays activation, and prototype logits are blended with segmentation logits (weight 0.1) during inference.

3.5 *Boundary-aware Module*

The Boundary-aware module (BA) operates differently during training and inference. During training, boundary points are identified from ground-truth labels as locations where neighboring points carry different class labels. These points receive enhanced features through learned boundary-specific weights, and an auxiliary binary head learns to predict boundary membership. A scale factor of 2.0 up-weights boundary points in the segmentation loss. During inference, the trained auxiliary head generates soft boundary probability estimates from input features alone, modulating the feature representations to focus on likely transition regions without requiring ground-truth labels.

3.6 *Composite Loss Function*

The total loss combines five objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} + \lambda_{\text{lov}}\mathcal{L}_{\text{Lovász}} + \lambda_{\text{bnd}}\mathcal{L}_{\text{boundary}} + \lambda_{\text{con}}\mathcal{L}_{\text{contrastive}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} \quad (2)$$

where $\mathcal{L}_{\text{CE}}$ is cross-entropy with label smoothing 0.1, $\mathcal{L}_{\text{Lovász}}$ (Berman et al. 2018) optimizes IoU directly, $\mathcal{L}_{\text{boundary}}$ is the boundary-weighted loss, $\mathcal{L}_{\text{contrastive}}$ is the contrastive loss, and $\mathcal{L}_{\text{aux}}$ is auxiliary boundary detection. Weights: $\lambda_{\text{CE}} = \lambda_{\text{lov}} = 1.0$, $\lambda_{\text{bnd}} = 0.3$, $\lambda_{\text{con}} = 0.05$, and $\lambda_{\text{aux}} = 0.1$..

4 Experiments

4.1 Dataset and Implementation

All experiments use the Saitama dataset, a LiDAR point cloud dataset for urban infrastructure monitoring with 19 scan areas annotated across six classes, as shown in Figure 1: Outdoor Ground (0), Vegetation (1), Building (2), Power Lines (3), Vehicle (4), and Indoor Ground Surface (5). The dataset has previously been used for transformer-based classification of urban infrastructure scenes, demonstrating its suitability for semantic interpretation in complex urban environments (Itakura et al. 2025). Areas 1–15 are used for training, 16–17 for validation, and 18–19 for testing. In the present study, particular attention is given to the distinction between Outdoor Ground and Indoor Ground Surface, since these two classes exhibit similar geometry but differ in material and contextual characteristics, while also showing a highly imbalanced class distribution.

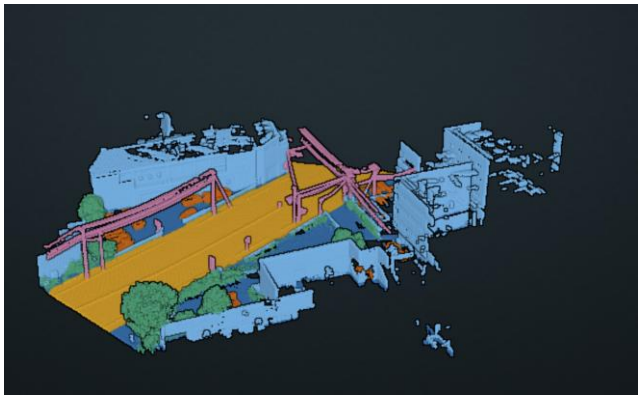

Figure 1. Class representation of dataset sample

Models are implemented in the Pointcept framework. Input clouds are voxelized at 0.04 m, with augmentation including random rotation (±1°), scaling (0.9–1.1×), flipping, jitter ($\sigma = 0.005$), chromatic augmentation, and sphere cropping to 40,960 points. Training uses batch size 4, AdamW (weight decay 0.05, gradient clipping 1.0), and

OneCycleLR with 5% warmup over 100 epochs. For transfer learning, the backbone uses 5×10^{-5} and the head uses 5×10^{-4}. Test-time augmentation with six variants is applied. The primary metric is mean Intersection-over-Union (mIoU).

4.2 Ablation Study

Seven configurations (B0–B6), all from the same nuScenes pre-trained backbone, evaluate the three enhancement modules. Table 1 reports the results.

Table 1. Ablation study results. MSCA: Multi-Scale Context Aggregation; SCL: Supervised Contrastive Learning; BA: Boundary-aware. Best values in bold.

Config	MSCA	SCL	BA	mIoU	Road	Veg.	Build.	Power	Vehicle	Indoor
B0 (Base)				0.8216	0.7798	0.8687	0.9264	0.9313	0.8421	0.5811
B1			✓	0.8249	0.8080	0.8586	0.9240	0.9360	0.8342	0.5888
B2		✓		0.8061	0.7193	0.8731	0.9240	0.9470	0.8248	0.5486
B3	✓			0.8224	0.7775	0.8682	0.9291	0.9387	0.8399	0.5810
B5	✓	✓		0.8335	0.8327	0.8686	0.9259	0.9495	0.8202	0.6043
B6		✓	✓	0.8277	0.8331	0.8718	0.9233	0.9434	0.8101	0.5844
B4 (Full)	✓	✓	✓	0.8368	0.8023	0.8772	0.9325	0.9525	0.8503	0.6061

Among individual modules, the BA module provides the most consistent improvement (B1: +0.0033 mIoU; Road 0.7798→0.8080). Contrastive learning alone (B2) reduces mIoU by 0.0155, likely because prototype optimization requires complementary modules for difficult classes (Indoor Ground drops from 0.5811 to 0.5486). Paired combinations outperform single modules: Context+Contrastive (B5: 0.8335) achieves the highest Road IoU at 0.8327. The full model (B4: 0.8368) achieves the best overall mIoU and leads in five of six classes, with Indoor Ground improving by 4.3% (0.5811→0.6061).

4.3 Transfer Learning Effectiveness

Table 2 compares the baseline, full ablation model, and final fine-tuned model.

Table 2. Quantitative comparison of the baseline model, full ablation model, and final fine-tuned model.

Method	mIoU	Road	Veg.	Build.	Power	Vehicle	Indoor
B0 (Baseline)	0.8216	0.7798	0.8687	0.9264	0.9313	0.8421	0.5811
B4(Full model)	0.8368	0.8023	0.8772	0.9325	0.9525	0.8503	0.6061
Fine-tuned (Ours)	0.8471	0.8308	0.8787	0.9315	0.9412	0.8681	0.6325

The fine-tuned model achieves the highest mIoU of 0.8471 (+2.55% over baseline, +1.03% over B4). The Indoor Ground Surface class improves by 5.14% points (0.5811→0.6325), confirming that transfer learning and the proposed enhancements are jointly most beneficial for underrepresented and ambiguous surface classes.

Figure 2. Per-class IoU heatmap across ablation configurations. Darker shading indicates higher IoU. The B4 (Full) row is outlined in orange.

4.4 Qualitative Results

Figure 3 provides a qualitative comparison at the boundary between Outdoor Ground and Indoor Ground Surface. The baseline model produces fragmented predictions near the surface transition, whereas the fine-tuned model yields cleaner and more spatially coherent segmentation, corroborating the quantitative improvements.

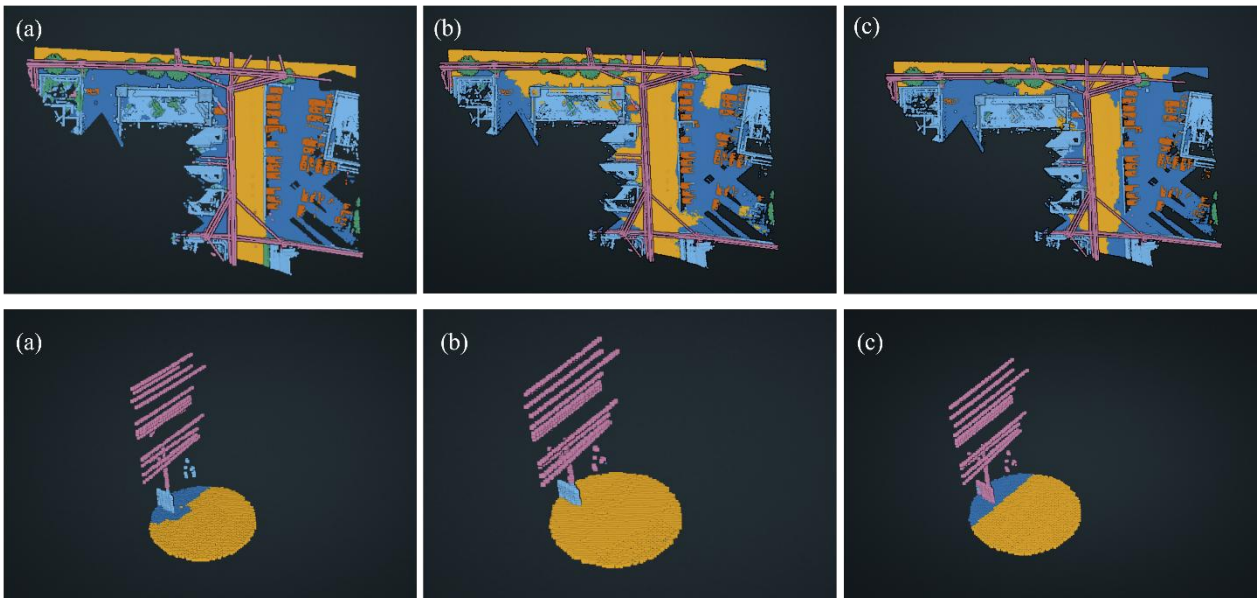

Figure 3. Qualitative comparison at a boundary region between Outdoor Ground (orange) and Indoor Ground Surface (blue). Left: Ground Truth. Center: Baseline prediction. Right: Fine-tuned prediction.

5 Conclusion

The present study proposes a PTv3-based model, a transfer-learning framework for the semantic segmentation of 3D LiDAR point clouds for civil engineering applications. The proposed method focuses on the segmentation of six-class urban infrastructure scenes. Emphasis is placed on the reliable separation of Outdoor Ground and Indoor Ground Surface. This challenging class pair shares similar local geometry but differs in broader spatial context and environmental characteristics.

To address this challenge, three task-specific modules are incorporated into a Point Transformer V3 (PTv3) backbone: MSCA for multi-scale contextual reasoning, SCL for improving feature discriminability between confusing classes, and BA for strengthening segmentation near ambiguous boundaries. Built on an initialization pre-trained on the nuScenes dataset, the framework combines large-scale transferable geometric priors with task-oriented architectural enhancements.

Experiments on the Saitama dataset demonstrate the effectiveness of the proposed framework. The final fine-tuned model achieves a validation mean Intersection-over-Union of 0.8471, improving upon the baseline by 2.55 percentage points and the full ablation model by 1.03 percentage points. The largest class-level gain occurs in the Indoor Ground Surface category, with the Intersection over Union increasing from 0.5811 to 0.6325. Furthermore, qualitative results show cleaner and more spatially coherent predictions near the boundary between Outdoor Ground and Indoor Ground Surface.

Overall, our results indicate that combining large-scale pre-training with task-specific enhancements is an effective strategy for point-cloud segmentation in the context of civil engineering, especially when labeled target data are limited and class ambiguity is pronounced. Future work will extend the framework to more diverse infrastructure datasets containing richer ground-related categories and will explore temporal or multi-session LiDAR data for more robust long-term infrastructure monitoring.

References

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C. and Gall, J., SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences, Proc. IEEE/CVF ICCV, 9297–9307, 2019.

Berman, M., Triki, A. R. and Blaschko, M. B., The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-over-Union Measure in Neural Networks, Proc. IEEE/CVF CVPR, 2018.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G. and Beijbom, O., nuScenes: A Multimodal Dataset for Autonomous Driving, Proc. IEEE/CVF CVPR, 11621–11631, 2020.

Chen, Y., Lin, C., Li, X., Kubo, S., Yamane, T. and Chun, P.-j., Automated Dimension Estimation of Bridge Components Using Semantic Segmentation and Geometric Fitting of Point Cloud Data, Eng. Struct., 342, 120837, 2025.

Choy, C., Gwak, J. and Savarese, S., 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks, Proc. IEEE/CVF CVPR, 3075–3084, 2019.

Fang, Z. et al., Advancing 3D Point Cloud Understanding through Deep Transfer Learning: A Comprehensive Survey, Inform. Fusion, 2024.

Itakura, K., Aruga, K. and Chun, P.-j, Exploring Transformer-Based Classification Methods for 3D Point Clouds of Urban Infrastructure, Artif. Intell. Data Sci., 6(3), 692–702, 2025.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C. and Krishnan, D., Supervised Contrastive Learning, Advances in Neural Inform. Process. Syst. (NeurIPS), 2020.

Lin, C., Abe, S., Zheng, S., Li, X. and Chun, P.-j., A Structure-Oriented Loss Function for Automated Semantic Segmentation of Bridge Point Clouds, Comput.-Aided Civ. Infrastruct. Eng., 40(6), 801–816, 2025.

Lin, C., Chen, Y., Itakura, K., Maharjan, S. and Chun, P.-j., Bridge Inspection Using Image–Point Cloud Fusion with Image Filtering, Damage Detection and 3D Registration, Autom. Constr., 180, 106538, 2025.

Qi, C. R., Su, H., Mo, K. and Guibas, L. J., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, Proc. IEEE/CVF CVPR, 652–660, 2017a.

Qi, C. R., Yi, L., Su, H. and Guibas, L. J., PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space, Advances in Neural Inform. Process. Syst. (NeurIPS), 2017b.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F. and Guibas, L. J., KPConv: Flexible and Deformable Convolution for Point Clouds, Proc. IEEE/CVF ICCV, 6411–6420, 2019.

Wu, X., Lao, Y., Jiang, L., Liu, X. and Zhao, H., Point Transformer V2: Grouped Vector Attention and Partition-based Pooling, Advances in Neural Inform. Process. Syst. (NeurIPS), 2022.

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T. and Zhao, H., Point Transformer V3: Simpler, Faster, Stronger, Proc. IEEE/CVF CVPR, 2024.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M. and Solomon, J. M., Dynamic Graph CNN for Learning on Point Clouds, ACM Trans. Graph., 38(5), 146, 2019.

Zhao, H., Jiang, L., Jia, J., Torr, P. H. S. and Koltun, V., Point Transformer, Proc. IEEE/CVF ICCV, 16259–16268, 2021.