Computer Vision and Image Understanding 262 (2025) 104531

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

SCENE-Net: Geometric induction for interpretable and low-resource 3D pole detection with Group-Equivariant Non-Expansive Operators

Diogo Lavado [a,b,*], Alessandra Micheletti [b], Giovanni Bocchi [b], Patrizio Frosini [c], Cláudia Soares [a]

[a] NOVA School of Science and Technology, Portugal
[b] University of Milan, Italy
[c] University of Bologna, Italy

ARTICLE INFO

Communicated by Juergen Gall

MSC:
68T45
58K70

Keywords:
3D semantic segmentation
Point clouds
Group Equivariant Non-Expansive Operators
White-box models
Power grids

ABSTRACT

This paper introduces SCENE-Net, a novel low-resource, white-box model that serves as a compelling proof-of-concept for 3D point cloud segmentation. At its core, SCENE-Net employs Group Equivariant Non-Expansive Operators (GENEOs), a mechanism that leverages geometric priors for enhanced object identification. Our contribution extends the theoretical landscape of geometric learning, highlighting the utility of geometric observers as intrinsic biases in analyzing 3D environments. Through empirical testing and efficiency analysis, we demonstrate the performance of SCENE-Net in detecting power line supporting towers, a key application in forest fire prevention. Our results showcase the superior accuracy and resilience of our model to label noise, achieved with minimal computational resources—this instantiation of SCENE-Net has only eleven trainable parameters—thereby marking a significant step forward in trustworthy machine learning applied to 3D scene understanding. Our code is available in: https://github.com/dlavado/scene-net.

1. Introduction

Explainable AI (XAI) methodologies have been rising in prominence due to the ever wider application field of deep learning, now including high-risk tasks such as autonomous driving and medical image processing. Most methods in XAI provide *post hoc* explanations to black-box models (i.e., algorithms unintelligible to humans). However, these are often limited in terms of their model fidelity (Lipton, 2018; Rudin, 2019), that is, they provide possible explanations subject to human interpretation for the predictions of the underlying model (e.g., heatmaps (Chefer et al., 2021; Voita et al., 2019) and input masks (Ribeiro et al., 2016; Fong and Vedaldi, 2017)), instead of providing a mechanistic understanding of their inner-workings. Conversely, intrinsic interpretability methods (i.e., white-box models) provide an understanding of their decisions through their architecture and parameters (Rudin, 2019). Transparency is achieved by enforcing constraints that reflect domain knowledge and simple designs (Lipton, 2018; Chen et al., 2020), which may impose a trade-off between performance and interpretability. Explainability in point clouds is a less explored area of research when compared to image data, most methods offer possible explanations for model predictions and are based on previously established 2D techniques, such as saliency maps (Zheng et al., 2019), LIME (Tan and Kotthaus, 2022), gradient-based explanations (Sundararajan et al., 2017; Matrone et al., 2022; Levi and Gilboa, 2024).

To the best of our knowledge, there are no white-box methods specifically designed for 3D point clouds. Thus, in this paper we propose a novel white-box model, **SCENE-Net**, that provides intrinsic geometric interpretability by leveraging on group equivariant non-expansive operators (GENEOs) (Bergomi et al., 2019; Cascarano et al., 2021; Bocchi et al., 2025a,b). GENEOs serve as functional observers (i.e., operators that signal the detection of patterns of interest) parameterized by meaningful geometric features. We define signature shapes found in point clouds to encode relevant properties and utilize them to construct intricate observers for detecting 3D elements. Our method goes beyond the straightforward template matching approach (Hashemi et al., 2016). Instead of relying on predefined shapes, the parameters defining our observers are learned from the data through backpropagation. Moreover, their convex combination facilitates the emergence of complex observers directly from data. SCENE-Net stands as a compelling proof-of-concept, highlighting the effectiveness of GENEOs in addressing purely geometric challenges without the reliance on color or additional features. Our model offers inherent interpretability, resilience to noisy labeling, and resource efficiency, with a streamlined

* Corresponding author at: University of Milan, Italy.
E-mail address: d.lavado@campus.fct.unl.pt (D. Lavado).

https://doi.org/10.1016/j.cviu.2025.104531
Received 11 September 2024; Received in revised form 30 September 2025; Accepted 9 October 2025
Available online 27 October 2025

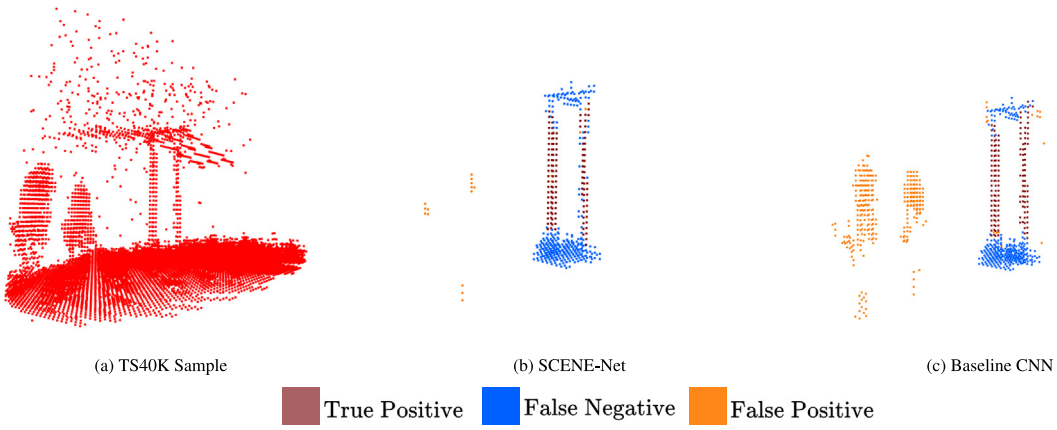

Fig. 1. Signature shapes for power line supporting tower detection. For our TS40K sample shown in (a), SCENE-Net accurately detects the body of the tower (b), while a comparable CNN has a large false positive area in the vegetation (c). Our model is interpretable with 11 trainable geometric parameters whereas the CNN has a total of 2190 parameters. The ground and power lines are mislabeled in the ground truth.

architecture of just 11 trainable parameters. In summary, we study the following research questions: **(1) How can we interpret SCENE-Net and its prediction?** We elucidate the intrinsic meaning of the eleven meaningful shape parameters and provide *post hoc* analyses of specific predictions that are independent from human interpretation. (Section 5.2.1). **(2) How does the performance of SCENE-Net fare against other methods?** In addition to comparing performance against a baseline CNN, we assess our model in the *SemanticKITTI* benchmark (Section 5.2.2). **(3) How effectively can we learn the geometry of 3D point clouds with SCENE-Net?** We showcase the complex observer learned through the composition of simple 3D shapes and demonstrate and study ablated model of our architecture (Section 5.2.3).

2. Related work

2.1. Point cloud segmentation

Processing point clouds for semantic segmentation presents unique challenges, largely due to their inherent unstructured nature and permutation invariance. Voxel-based strategies have been a primary approach to endow point clouds with a lattice structure, enabling the application of 3D convolutions (Long et al., 2015; Rethage et al., 2018; Tchapmi et al., 2017). However, these methods have significant drawbacks, such as the excessive memory requirements for high-resolution voxel grids. To address these limitations, researchers have explored alternative strategies like sparse convolutions (Graham et al., 2018; Su et al., 2018) and octree-based CNNs (Wang et al., 2017; Le and Duan, 2018) to efficiently process point clouds. Directly processing point clouds has also gained traction, inspired by the seminal work of PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b). These efforts laid the foundation for using point sub-sampling and feature aggregation techniques to extract local features effectively. Subsequent convolution-based methods (Hua et al., 2018; Li et al., 2018; Wu et al., 2019; Thomas et al., 2019; Zhu et al., 2021) have shown promising results. But recent advancements in transformer-based architectures are achieving state-of-the-art performance on all key benchmarks (Zhao et al., 2021; Wu et al., 2022, 2024; Wang, 2023; Liu et al., 2023b; Lai et al., 2023).

SCENE-Net is a voxel-based architecture that introduces a time-efficient solution for processing high-resolution grids, with support for sizes from 64^3 to 256^3, addressing a crucial gap in the current landscape. Furthermore, our model can be used as an efficient and robust geometric feature extraction strategy for black-box models. Instead of just relying on the spatial coordinates of 3D points, SCENE-Net can be used as a pre-processing step and trained simultaneously with black-box models.

2.2. Explainability on 3D point clouds

Explainability is essential in ML, especially in high-stakes applications where understanding model decisions is crucial, as explored in Lipton (2018), Guidotti et al. (2018) and Doshi-Velez and Kim (2017). The literature predominantly categorizes explainability approaches into *post hoc* explainability and intrinsic interpretability. *Post hoc* techniques, such as LIME (Ribeiro et al., 2016), meaningful perturbations (Fong and Vedaldi, 2017, anchors Ribeiro et al., 2018), and ontologies (Ribeiro and Leite, 2021), offer explanations for black-box models' predictions by correlating inputs with outputs. Despite their flexibility and model-agnostic nature, these methods often miss deeper cause–effect insights and can misconstrue feature significance, sometimes equating the explanatory power of heterogeneous inputs like a dog image and random noise similarly (Rudin, 2019). They also add computational overhead, challenging their utility in complex tasks such as 3D semantic segmentation (Zheng et al., 2019; Tan and Kotthaus, 2022; Sundararajan et al., 2017; Matrone et al., 2022; Levi and Gilboa, 2024). In contrast, intrinsic interpretability embeds explanations within the architecture of the model, as exemplified by decision trees and linear models. These white-box models traditionally trade-off complexity for transparency, limiting their depth in favor of simplicity and domain-specific constraints, as argued in Rudin (2019). However, recent techniques like concept whitening (Chen et al., 2020) and interpretable CNNs (Zhang et al., 2018) demonstrate that interpretability need not sacrifice performance, offering insights directly tied to the inner workings of the model, albeit still dependent on human-centric interpretations. SCENE-Net embeds intrinsic geometric interpretability directly into its architecture, bypassing the need for human interpretation biases. By leveraging prior geometric knowledge and encoding it through functional observers whose parameters are learned during training, SCENE-Net provides mechanistic insights into its decision-making process. This approach not only ensures high-level mathematical simplicity but also capitalizes on the power of convolutional kernels, presenting a significant advance in creating interpretable, efficient models for complex tasks like 3D semantic segmentation.

2.3. Group equivariant methodologies

Group equivariant techniques have been explored in computer vision with group convolutions (G-convolutions) (Cohen and Welling, 2016; Cohen et al., 2019). This work focuses on generalizing the conventional translation-equivariant convolution to arbitrary groups via physically parameterized kernel functions. For example, (Cohen and Welling, 2016) illustrate G-convolutions using the *p4* group, incorporating translations and 90-degree rotations, and the *p4m* group, which also

includes reflections. In practice, *p4*-convolutional kernels are rotated in 90, 180 and 270 degrees and stacked to then convolve the input data. The theoretical framework on GENEOs (Bergomi et al., 2019; Cascarano et al., 2021) is a generalization of G-convolutions (Cohen and Welling, 2016; Cohen et al., 2019): (Bergomi et al., 2019) define general group equivariant operators that transform a topological space into another while respecting a group of transformations G. This definition can be instantiated with the convolution operator and applied to the Euclidean space (a specific instance of a topological space) where G-convolutions (Cohen and Welling, 2016) are defined. In essence, the convolution operator is a GENEO. Bergomi et al. (2019) highlight that limiting to specific operator families and maintaining equivariance to interpretable transformations (e.g., translation) are crucial for the success of CNNs. Unlike Cohen et al. (2019), the GENEO framework imposes a non-expansiveness condition on GENEO-kernels to ensure faster convergence and approximation by a finite set of operators within the same space (Bergomi et al., 2019; Cascarano et al., 2021). This is not an essential requirement for defining general G-equivariant operators, as Group Equivariant Operators (GEOs) are also included within the GENEO framework (Bergomi et al., 2019; Cascarano et al., 2021).

The GENEO-kernels in SCENE-Net are based on the convolution operator to take advantage of its equivariance w.r.t. translation, a crucial property for feature extraction in 3D point clouds. However, our kernels do not align with the definition of G-convolutions in Cohen and Welling (2016): the latter generalizes the shift in the conventional convolution operator to encompass other types of isometries, whereas our GENEO-kernels achieve equivariance to additional groups with kernels parameterized by geometric meaningful features. This way, we are not limited to isometries and achieve transparency to the user in our feature extraction process since our meaningful parameters and learned from data.

3. Group equivariant non-expansive operators

GENEOs are the building blocks of a mathematical framework (Bergomi et al., 2019) that formally describes machine learning agents as a set of operators acting on the input data. These operators provide a measure of the world, just as CNN kernels learn essential features to, for instance, recognize objects. Such agents can be thought of as observers that analyze data. They transform it into higher-level representations while respecting a set of properties (i.e., a group of transformations). An appropriate observer transforms data in such a way that respects the right group of transformations, that is, it commutes with these transformations. Formally, we say that the observer is *equivariant* with respect to a group of transformations. The framework takes advantage of topological data analysis (TDA) to describe data as topological spaces. Specifically, a set of data X is represented by a topological space Φ with admissible functions $\varphi : X \to \mathbb{R}^3$. Φ can be thought of as a set of admissible measurements that we can perform on the measurement space X. For example, images can be seen as functions assigning RGB values to pixels. This not only provides uniformity to the framework but also allows us to shift our attention from raw data to the space of measurements that characterizes it. Now that the input data is well represented, let us introduce how the framework defines prior knowledge. Data properties are defined through maps from X to X that are Φ-preserving homeomorphisms. That is, the composition of functions in Φ with such homeomorphisms produces functions that still belong to Φ. Therefore, we can define a group G of Φ-preserving homeomorphisms, representing a group of transformations on the input data for which we require equivariance to be respected. In other words, G is the group of properties that we chose to enforce equivariance w.r.t. the geometry in the original data. It is through G that we embed prior knowledge into a GENEO model. Following the previous example, planar translations can define a subgroup of G.

Let us consider the notion of a *perception pair* (Φ, G): it is composed of all admissible measurements Φ and a subgroup of Φ-preserving homeomorphisms G.

Definition 1 (*Group Equivariant Non-Expansive Operator (GENEO)*)**.** Consider two perception pairs (Φ, G) and (Ψ, H) and a homomorphism $T : G \to H$. A map $F : \Phi \to \Psi$ is a group equivariant non-expansive operator if it exhibits equivariance:

$$\forall \varphi \in \Phi, \forall g \in G, F(\varphi \circ g) = F(\varphi) \circ T(g) \tag{1}$$

and is non-expansive:

$$\forall \varphi_1, \varphi_2 \in \Phi, \|F(\varphi_1) - F(\varphi_2)\|_\infty \leq \|\varphi_1 - \varphi_2\|_\infty \tag{2}$$

Non-expansivity and convexity are essential for the applicability of GENEOs in a machine-learning context. When the spaces Φ and Ψ are compact, non-expansivity guarantees that the space of all GENEOs $\mathcal{F}$ is compact as well. Compactness ensures that any operator can be approximated by a finite set of operators sampled in the same space. Moreover, by assuming that Ψ is convex, Bergomi et al. (2019) proves that $\mathcal{F}$ is also convex. These properties are essential for the applicability of GENEOs in a machine learning paradigm. Convexity guarantees that the convex combination of GENEOs is also a GENEO. Therefore, these results prove that any GENEO can be efficiently approximated by a certain number of other GENEOs in the same space.

In addition to drastically reducing the number of parameters in the modeling of the considered problems and making their solution more transparent, we underline that the use of GENEOs makes available various theoretical results that allow us to take advantage of a new mathematical theory of knowledge engineering. Lastly, besides the cited compactness and convexity theorems, algebraic methods concerning the construction of GENEOs are already available (Bocchi et al., 2022; Conti et al., 2022; Botteghi et al., 2020; Bocchi et al., 2025a), as well as results on the robustness of the approach (Bocchi et al., 2025b).

3.1. GENEOs as inductive biases in 3D scene understanding

Inductive biases remain crucial for effective research in machine learning. The no-free-lunch theorem emphasizes the necessity of these biases, asserting that no single learning algorithm excels at all possible tasks without incorporating specific assumptions or constraints (Baxter, 2000; Goyal and Bengio, 2022). While general methods leveraging massive computation often outperform those incorporating domain-specific knowledge, it is important to acknowledge the effective use of inductive biases. Lavie et al. (2024) highlight that transformers, despite their general-purpose nature, embody strong inductive biases, such as a tendency toward more permutation symmetric functions in sequence space. This is elucidated through the representation theory of the symmetric group, which aids in understanding and predicting the learnability of functions under permutation symmetry. Similarly, our approach with SCENE-Net employs geometric priors through GENEOs, embedding intrinsic interpretability and efficiency into the model architecture. These inductive biases are vital for tasks involving 3D point clouds, particularly in resource-constrained environments like drone-based applications. By encoding domain-specific geometric knowledge directly into the model, SCENE-Net achieves high accuracy and resilience to label noise with minimal computational resources. This approach aligns with the practical needs of real-world deployment and the long-term trends favoring scalable computation. Thus, the integration of carefully chosen inductive biases, such as those provided by GENEOs, enhances both performance and interpretability. This reinforces the ongoing relevance of inductive biases in advancing research and practical applications in machine learning.

4. SCENE-Net: Signature geometric equivariant non-expansive operator network

In this section, we introduce the overall architecture of **SCENE-Net**. Next, we define the geometrical properties that describe power line supporting towers. Lastly, we detail the loss function used to train the observer.

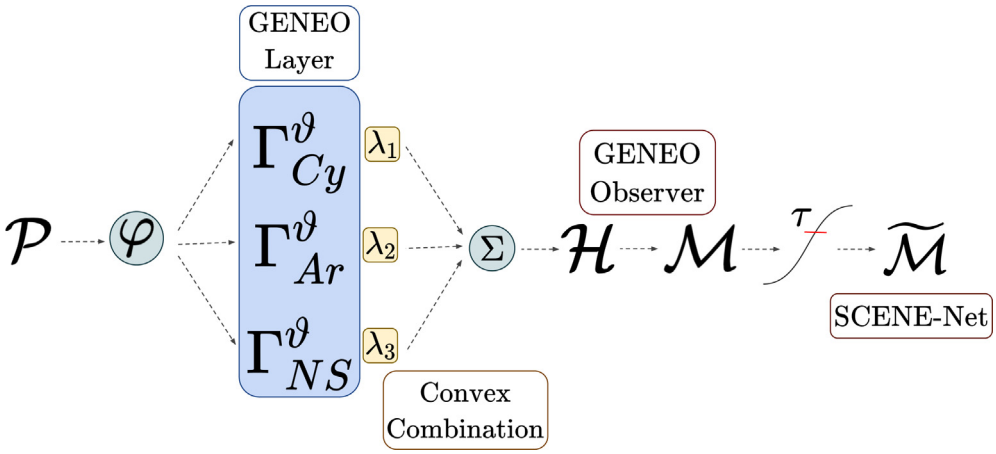

Fig. 2. Pipeline of SCENE-Net: an input point cloud $\mathcal{P}$ is measured according to function φ and voxelized. This representation then is fed to a GENEO-layer, where each operator $\Gamma_i^{\vartheta_i}$ separately convolves the input. A GENEO observer $\mathcal{H}$ is then achieved by a convex combination of the operators in the GENEO layer. $\mathcal{M}$ transforms the analysis of the observer into a probability of belonging to a tower. Lastly, a threshold operation is applied to classify the voxels. Note that this final step occurs after training is completed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1. Overview

3D Point clouds are generally denoted as $\mathcal{P} \in \mathbb{R}^{N\times(3+d)}$, where N is the number of points and $3+d$ is the cardinality of spatial coordinates plus any point-wise features, such as colors or normal vectors. The input point cloud is first transformed in accordance with a measurement function $\varphi : \mathbb{R}^3 \to \{0,1\}$ that signals the presence of 3D points in a voxel discretization. Next, the transformed input is fed to a layer of multiple GENEOs (GENEO-layer), each chosen randomly from a parametric family of operators, and defined by a set of trainable shape parameters ϑ (Fig. 2). Such GENEOs are in the form of convolutional operators with carefully designed kernels as described later. Not only is convolution a well-studied operation, but it also offers equivariance w.r.t. translations by definition. During training, it is not the kernels themselves that are fine-tuned with back-propagation, since this would not preserve equivariance at each optimization step. Instead, the error is propagated to the shape parameters ϑ of each operator. Following the GENEO-layer, its set of operators $\Gamma = \{\Gamma_i^{\vartheta_i}\}_{i=1}^K$, with shape parameters $\vartheta = \vartheta_1, \dots, \vartheta_k$, are combined through convex combination with weights $\lambda = (\lambda_1, \dots, \lambda_k)^T$ by $\underset{\lambda,\vartheta}{\mathcal{H}} : \mathcal{P} \to \mathcal{P}$ such that

$$\underset{\lambda,\vartheta}{\mathcal{H}}(x) = \sum_{i=1}^{K} \lambda_i \Gamma_i^{\vartheta_i}(\varphi)(x) \tag{3}$$

Since the convex combination of GENEOs is also a GENEO (Bergomi et al., 2019), $\mathcal{H}$ preserves the equivariance of each operator $\Gamma^{\vartheta} \in \Gamma$. In fact, $\mathcal{H}$ defines a GENEO observer that analyzes the 3D input scenes looking for the geometrical properties encoded in Γ. The convex coefficients λ represent the overall contribution of each operator Γ_i^{ϑ} to the $\mathcal{H}$ analysis. The parameters grant our model its intrinsic interpretability. They are learned during training and represent geometric properties and the importance of each Γ^{ϑ} in modeling the ground truth.

Next, we transform the observer's analysis into a probability of each 3D voxel belonging to a supporting tower as a model $\underset{\lambda,\vartheta}{\mathcal{M}} : \mathcal{P} \to [0,1]^N$

$$\underset{\lambda,\vartheta}{\mathcal{M}}(x) = \left(\tanh\left(\underset{\lambda,\vartheta}{\mathcal{H}}(x)\right)\right)_+ \quad , \tag{4}$$

where $(t)_+ = \max\{0, t\}$ is the rectified linear unit (ReLU).

The output of $\mathcal{H}$ is geometrically meaningful: it is a real-valued field where positive regions indicate strong alignment with pole-like structures, and negative regions signal inconsistency with the sought-out patterns. To convert this into a semantically valid probability distribution, we first compress the output using tanh, bounding values to $[-1,1]$ and dampening the magnitude of outliers. We then apply ReLU to suppress all negative responses, which conversely entails that only positively aligned structures are assigned a non-zero probability. This combination produces a probability map $\mathcal{M}$ bounded in $[0,1]$ with interpretable semantics illustrated in Fig. 8: the activation of each voxel reflects the strength of pole-shaped geometry in that region. We note that sigmoid was not used because it would assign non-zero probability to negatively aligned voxels, contradicting the semantics of $\mathcal{H}$. Similarly, using ReLU alone without prior normalization would not ensure bounded or calibrated outputs.

Lastly, a probability threshold $\tau \in [0,1]$ is defined through hyperparameter fine-tuning and applied to $\mathcal{M}$ resulting in a map $\underset{\lambda,\vartheta}{\widetilde{\mathcal{M}}} : \mathcal{P} \times \mathbb{R} \to \{0,1\}^N$

$$\underset{\lambda,\vartheta}{\widetilde{\mathcal{M}}}(x,\tau) = \left\{\underset{\lambda,\vartheta}{\mathcal{M}}(x)\right\} \geq \tau \quad , \tag{5}$$

where $\widetilde{\mathcal{M}}$ denotes the **SCENE-Net** model.

4.2. Encoding geometric priors with GENEOs for power line tower identification

In this section, we formalize the observer $\mathcal{H}$, specifically tailored to distinguish power line supporting towers within their environments.

4.2.1. Cylinder GENEO

These GENEOs will be equivariant with respect to the group of all isometries of $\mathbb{R}^3$ that map upward-oriented vertical lines to upward-oriented vertical lines and preserve the orientation of $\mathbb{R}^3$, when $T : G \to G$ is the identity homomorphism. Thus, the distinctive verticality of towers amidst rural landscapes is captured by cylinder GENEOs. This group G includes translations in the xy plane and rotations around the z-axis.

Definition 2. A smoothed cylinder pattern $g_{Cy} : \mathbb{R}^3 \to [0,1]$ is given by:

$$g_{Cy}(x) = e^{-\frac{1}{2\sigma^2}(\|\pi_{-3}(x)-\pi_{-3}(c)\|^2 - r^2)^2}$$

where $\pi_{-3}(x)$ projects x onto the xy plane, and c denotes the center of the cylinder. The parameters $\vartheta_{Cy} = [r, \sigma]$ define the radius of the cylinder and the spread of the smoothing Gaussian.

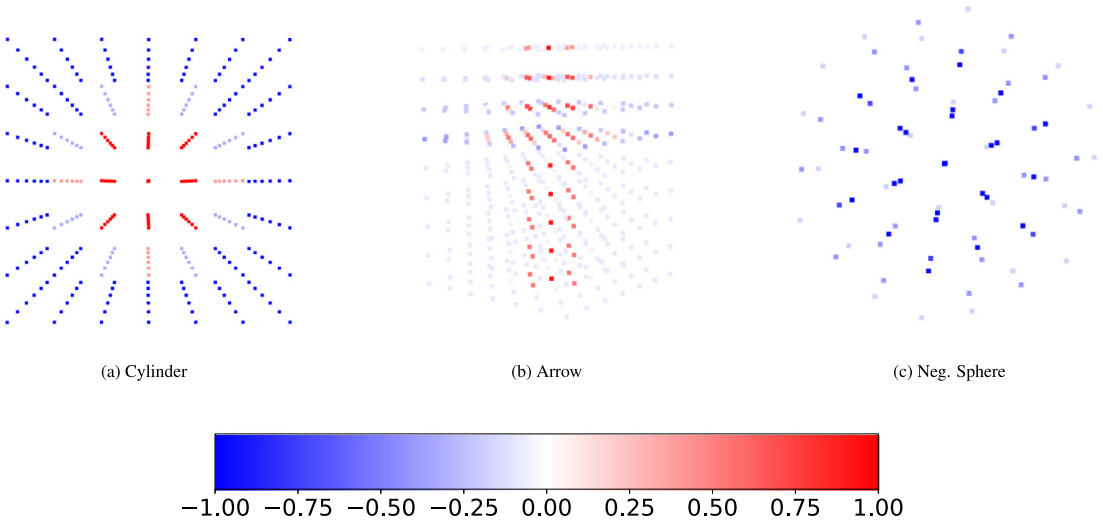

Fig. 3. GENEO kernels discretized in a voxel grid and colored according to weight distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

GENEOs act on functions, transforming them to remain equivariant to a specific group of transformations. Our GENEOs act upon Φ, the topological space representing $\mathcal{P}$ with admissible functions $\varphi : \mathbb{R}^3 \to \{0,1\}$. Specifically, we work with appropriate $\varphi \in \Phi$ functions that represent point clouds and preserve their geometry. For instance, φ can be a function that signals the presence of 3D points in a voxel grid. Therefore, the cylinder GENEO Γ^{ϑ}_{Cy} transforms φ into a new function that detects sections in the input point cloud that demonstrate the properties of g_{Cy} and, simultaneously, preserves the geometry of the 3D scene

$$\Gamma^{\vartheta}_{Cy} : \Phi \to \Psi, \qquad \psi_{Cy} = \Gamma^{\vartheta}_{Cy}(\varphi)$$
$$\psi_{Cy}(x) = \int_{\mathbb{R}^3} \tilde{g}_{Cy}(y)\varphi(x-y)dy \tag{6}$$

where Ψ is a new topological space that represents $\mathcal{P}$ with functions $\psi : \mathbb{R}^3 \to [0,1]$ and $\tilde{g}_{Cy}$ defines a normalized cylinder. The kernel g_{Cy} is normalized to have a zero-sum to promote the stability of the observer. This way, we encourage the geometrical properties that exhibit the sought-out group of transformations and punish those which do not. Thus, $\psi_{Cy}(x)$ assumes positive values for 3D points near the radius, whereas negative values discourage shapes that do not fall under the g_{Cy} definition. This leads to a more precise detection of the encoded group of transformations.

In other words, the cylinder GENEO Γ^{ϑ}_{Cy} acts on Φ to enhance cylindrical structures while preserving the geometry of the scene. Fig. 3(a) illustrates the kernel g_{Cy} discretized in a voxel grid.

Theorem 1. *The operator Γ^{ϑ}_{Cy} is non-expansive.*

Proof. Take the definition of our cylinder GENEO as

$$\int_0^h \int_{\mathbb{R}^2} g_{Cy}(x)dx = h \int_{\mathbb{R}^2} e^{-\frac{1}{2\sigma^2}(\|\tilde{x}-\tilde{c}\|^2 - r^2)^2} d\tilde{x} \tag{7}$$

where $\tilde{x}$ and $\tilde{c}$ are the projections on the first two coordinates. We start by a variable transformation $z = \tilde{x} - \tilde{c}$ yielding, as the inner integral

$$X = \int_{\mathbb{R}^2} e^{-\frac{1}{2\sigma^2}(\|\tilde{x}-\tilde{c}\|^2 - r^2)^2} dx = \int_{\mathbb{R}^2} e^{-\frac{1}{2\sigma^2}(\|z\|^2 - r^2)^2} dz. \tag{8}$$

We now consider the fact that there is a positive scalar $t^\star > 0$, such that for all $t > t^\star$, $t^2 - r^2 > t$. Thus, we break the integral in (8) in two terms, one integrating over the disk centered in zero with radius $t^\star$, $B(0,t^\star)$, and another integrating over the complementary region, outside the ball, $\bar{B}(0,t^\star)$

$$X = \underbrace{\int_{B(0,t^\star)} e^{-\frac{1}{2\sigma^2}(\|z\|^2 - r^2)^2} dz}_{X_0} + \int_{\bar{B}(0,t^\star)} e^{-\frac{1}{2\sigma^2}(\|z\|^2 - r^2)^2} dz.$$

The first integral, X_0, exists and is bounded by the area of the ball A_B multiplied by the maximum C_0 defined as

$$C_0 = \max_{B(0,t^\star)} e^{-\frac{1}{2\sigma^2}(\|z\|^2 - r^2)^2}.$$

Here, by the Weierstrass extreme value theorem, C_0 is a finite number because the function to be maximized is a continuous function over a compact set. The second integral can also be bounded. According to the definition of $t^\star$,

$$\|y\| > t^\star \implies \|y\|^2 - r^2 \geq \|y\| \implies e^{-\frac{(\|y\|^2 - r^2)^2}{2\sigma^2}} \leq e^{-\frac{\|y\|^2}{2\sigma^2}}.$$

the first inequality entails the second because $\|y\| \geq 0$.

The integral $C_1 = \int_{\bar{B}(0,t^\star)} e^{-\frac{\|y\|^2}{2\sigma^2}} dy$ can be easily computed. This entails that the overall integral is bounded by $C = h(A_B C_0 + C_1)$. The constant C can be used to normalize $g_{Cy}(x)$, making it non-expansive. □

4.2.2. Arrow GENEO

To distinguish towers from other vertical structures like trees, we introduce the Arrow GENEO. This operator combines a cylinder with a conical top, addressing the varied angles at which power lines intersect their supporting towers.

Definition 3. The Arrow observer $g_{Ar} : \mathbb{R}^3 \to [0,1]$ is defined as:

$$g_{Ar}(x) = \begin{cases} e^{\frac{-1}{2\sigma^2}(\|\pi_{-3}(x) - \pi_{-3}(c)\|^2 - r^2)^2} & \pi_3(x) < h \\ e^{\frac{-1}{2\sigma^2}(\|\pi_{-3}(x) - \pi_{-3}(c)\|^2 - (r_c \tan(\beta\pi))^2)^2} & \text{otherwise} \end{cases}$$

the radii of the cylinder and cone are defined by r and r_c, respectively, with c as their center. Lastly, h defines the height at which the cone is placed on top of the cylinder and $\beta \in [0, 0.5)$ defines the inclination of the cone.

Here, g_{Ar} characterizes a cone atop a cylinder, with shape parameters $\vartheta_{Ar} = [r, \sigma, h, r_c, \beta]$. Fig. 3(b) shows the discretization of the Arrow kernel.

Since the arrow is composed of a cylinder and a cone on top, the first part of the proof is identical, whereas, for the cone, we can assume an upper bound in the form of a cylinder with the same radius as the base of the cone and follow the same logic.

4.2.3. Negative sphere GENEO

To suppress spherical shapes common in vegetation, we employ a Negative Sphere GENEO. This operator is designed to penalize spherical geometries that could be mistaken for parts of trees and other arboreal structures.

Definition 4. The Negative Sphere $g_{NS} : \mathbb{R}^3 \to [-\omega, 1[$ is defined as

$$g_{NS}(x) = -\omega e^{\frac{-1}{2\sigma^2}(\|x-c\|^2 - r^2)^2}.$$

with $\omega \in]0, 1]$.

The role of the Negative Sphere is to de-emphasize non-target structures, with shape parameters $\vartheta_{NS} = [r, \sigma, \omega]$. Fig. 3(c) depicts the computation of this kernel in a voxel grid.

The proof of non-expansiveness for this kernel follows the same reasoning as the previous proofs. The main difference being that the sphere is not bounded by a height, so we integrate directly in $\mathbb{R}^3$ when defining the kernel.

4.3. Optimization GENEO embedding

The GENEO framework imposes a convex structure on the observer $\mathcal{H}$, which needs to be preserved during optimization. We formalize our learning objective as follows:

$$\begin{aligned} \underset{\lambda,\vartheta}{\text{minimize}} \quad & \mathbb{E}_{(X,y)\sim D}\left[\mathcal{L}_{seg}(\lambda, \vartheta; X, y)\right] \\ \text{s.t.} \quad & \lambda \in \Delta^{K-1}, \quad \vartheta \in \mathbb{R}_+^T, \end{aligned} \tag{9}$$

where Δ^{K-1} denotes the $(K-1)$-dimensional simplex, ensuring that the weights λ form a convex combination. $\mathbb{R}_+^T$ represents the non-negative orthant for shape parameters ϑ. Since the shape parameters ϑ of SCENE-Net are meaningful, they have a specific range where they hold their significance, for instance, the radius of a cylinder only holds meaning if it is non-negative.

The segmentation loss $\mathcal{L}_{seg}$ is defined as a weighted squared error term

$$\mathcal{L}_{seg}(\lambda, \vartheta; X, y) = f_w(\alpha, \epsilon; y)\Big(\underset{\lambda,\vartheta}{\mathcal{M}}(X) - y\Big)^2, \tag{10}$$

with $f_w(\alpha, \epsilon; y)$ denoting a weighting function that addresses class imbalance, parameterized by α and ϵ as detailed in Steininger et al. (2021). The expectation is taken over the data distribution D. The hyperparameter α emphasizes the weighting scheme, whereas ϵ is a small positive number that ensures positive weights. We use mean squared error (MSE) instead of binary cross-entropy (BCE) because SCENE-Net treats segmentation as a voxel-wise regression problem. Each voxel in the 3D space may partially intersect with a supporting tower; thus, we interpret the target as a continuous-valued density function, not a binary label. This reflects the physical nature of the data, where towers span multiple voxels and are not perfectly localized. To simplify supervision, we discretize tower occupancy as follows: if a voxel contains any tower point, its label is set to 1.0; otherwise, it is 0.0. In practice, we also found MSE to provide more stable training dynamics and better final performance than BCE in this regression-style setup.

To facilitate optimization, we reparameterize λ to satisfy the simplex constraint by setting $\lambda_K = 1 - \sum_{i=1}^{K-1} \lambda_i$. This reduces the problem to:

$$\begin{aligned} \underset{\lambda,\vartheta}{\text{minimize}} \quad & \mathbb{E}_{(X,y)\sim D}\left[\mathcal{L}_{seg}(\lambda, \vartheta; X, y)\right] \\ \text{s.t.} \quad & \vartheta \in \mathbb{R}_+^T, \quad \lambda \in \mathbb{R}_+^K, \end{aligned} \tag{11}$$

where we now only require non-negativity for λ and ϑ. We incorporate a regularization term to enforce non-negativity, leading to the final optimization problem

$$\underset{\lambda,\vartheta}{\text{minimize}} \quad \mathbb{E}_{(X,y)\sim D}\left[\mathcal{L}_{seg}(\lambda, \vartheta; X, y)\right] + \Omega(\lambda, \vartheta), \tag{12}$$

where $\Omega(\lambda, \vartheta) = \rho_l \sum_{i=1}^{K} h(\lambda_i) + \rho_t \sum_{i=1}^{K} \sum_{j=1}^{T_i} h(\vartheta_{ij})$ is the regularization term with $h(x) = \max(0, -x)$ and $\rho_l, \rho_t > 0$ are the regularization coefficients. The formulation in (12) strikes a balance between data fidelity, measured by the segmentation loss, and the theoretical guarantees of convexity and non-negativity required in the GENEO framework.

On the choice of constraining parameters to the simplex. A key component of SCENE-Net is the convex combination of GENEO kernels, where the weights λ are required to be non-negative. Instead of enforcing this constraint through softmax normalization or exponential reparameterization, we introduce a soft regularization penalty in Eq. (12) that discourages negative values while allowing the optimizer flexibility. Compared to the softmax function, our approach offers several advantages. Softmax forces the weights to sum to one and maps them to a probability simplex, but it does so by exponentiating the input vector. This introduces a bias toward peaked, imbalanced weight distributions: small differences between scores can result in large differences in final values. As a result, softmax tends to emphasize one dominant component and diminish others, which may hinder balanced contributions from multiple GENEOs. Similarly, exponential reparameterizations (e.g., $\lambda_i = \exp(\theta_i)$) guarantee positivity but also amplify disparities between components. This can make training unstable, obscure the contribution of each operator, and reduce model interpretability. Moreover, exponentiation eliminates the ability to impose structured regularization (e.g., sparsity or grouping) directly on the coefficients. In contrast, our approach treats λ as unconstrained parameters and biases them toward the non-negative orthant using a differentiable penalty. This allows the optimizer to find solutions that are both numerically stable and geometrically meaningful. In practice, we observe that all coefficients converge to non-negative values, and the regularization term vanishes at convergence, indicating that the constraint is satisfied without requiring explicit projection or reparameterization. Crucially, this formulation preserves interpretability: each weight in λ directly reflects the contribution of a specific GENEO kernel. It also enables structured control, such as applying L_1 regularization to encourage sparsity, something that is not straightforward with softmax or exponential transformations.

5. Experiments

5.1. TS40K dataset

We assess SCENE-Net using the TS40K dataset (Lavado et al., 2025), a distinctive collection of outdoor 3D point clouds spanning over 40,000 km of the European electrical transmission system in rural areas. This dataset plays a pivotal role in enhancing inspection procedures critical to averting power outages, grid damage, and forest fires. Leveraging drone-based LiDAR scans over conventional methods, TS40K offers unique 3D data characteristics: high point-density, minimal occlusion, and consistent point-density across regions, distinguishing it from self-driving benchmarks. This is crucial for geometric models such as SCENE-Net to harness the full potential of our geometric priors. Moreover, annotations based on inspections introduce noise and mislabeled points, mirroring real-world conditions, while a pronounced class imbalance presents a significant challenge, particularly with underrepresentation in classes related to the transmission system.

The simple and efficient architecture of SCENE-Net imposes a trade-off with domain application. Specifically, SCENE-Net is designed to detect pole-like structures, such as the power line supporting towers in the TS40K. In this context, detecting the towers lets us automatically derive the position of power lines. This enabled the use of SCENE-Net in an electrical utility company to accelerate inspection time and maintain the safety of power grids.

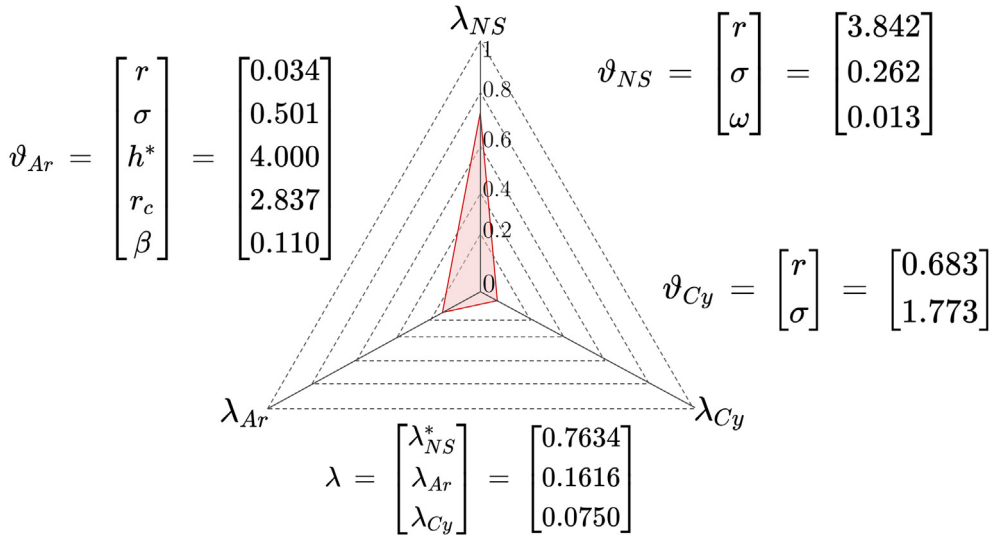

Fig. 4. The trainable parameters of SCENE-Net, ϑ and λ. Parameter h^* is not trainable, and λ^*_{NS} is defined as a function of the other mixing weights $\lambda^*_{NS} = 1 - \lambda_{Ar} - \lambda_{Cy}$.

5.2. Results and analysis

5.2.1. RQ1: The interpretability of SCENE-Net

The meaning of the learned shape parameters. To understand if the model parameters are interpretable, we inspect SCENE-Net's eleven trainable parameters ϑ and λ after training. Each $\vartheta_i \in \vartheta$ holds the learned shape parameters of a geometrical operator Γ_i, such as their height or radius. The convex coefficients λ weigh each operator Γ_i in our model's analysis. For example, we can conclude that the instance ϑ_{NS} of the Negative Sphere GENEO (Γ_{NS}) holds a weight of 76.34% on SCENE-Net's output (Fig. 4). The geometric nature of the observer and combination parameters endow intrinsic **interpretability** to SCENE-Net.

Post hoc interpretation for specific predictions. We can correlate the detection of scene elements, such as vegetation, to the contributions of each GENEO. This provides an extra layer of transparency to our model. The Arrow kernel is responsible for the detection of towers, the Cylinder aids this process and diminishes the detection of vegetation, and the Negative Sphere stabilizes the model by balancing contributions of the previous kernels (Fig. 5).

5.2.2. RQ2: The performance of SCENE-Net

Baseline comparison. We evaluate SCENE-Net on the task of identifying supporting towers in the TS40K dataset, using only geometric input with no color or auxiliary features. We compare SCENE-Net against a baseline CNN trained under the same protocol and using the same architecture: a single 3D convolutional layer with three kernels of size $9 \times 5 \times 5$, followed by a non-linear transformation, namely $\mathrm{ReLU}(\tanh(\cdot))$ for SCENE-Net and ReLU for the baseline CNN. The main difference lies in how these kernels are parameterized. The baseline CNN uses fully unconstrained convolution kernels with 2190 trainable parameters, whereas SCENE-Net restricts each kernel to a parametric form based on geometric priors, resulting in only 11 trainable parameters. In this sense, the baseline CNN has significantly more capacity and can, in principle, instantiate SCENE-Net if it finds the appropriate configuration in parameter space. Interestingly, the instantiation of kernel weights from parametric continuous functions makes SCENE-Net independent of kernel size. That is, the number of parameters of SCENE-Net is the same for any kernel size used to discretize its GENEO-kernels. This is of special importance in 3D applications, where memory and computational costs increase exponentially with the kernel size in traditional CNNs. We investigate this further in Section 5.2.3.

Table 1
3D semantic segmentation metrics on TS40K.

Method	Precision (%)	Recall (%)	IoU (%)
CNN	44 (± 7)	**26** (± 2)	53
SCENE-Net	**82** (± 8)	13 (± 5)	**58**

The application penalizes false positives more, thus we will emphasize Precision. Due to the imbalanced nature of the labels, we measured overall Precision, Recall, and Intersection over Union (IoU). Quantitatively, we observe a lift in Precision of 38%, and of 5% in IoU, and a drop of 13% in Recall (Table 1). In Figs. 1 and D.10, we can analyze the qualitative results on the TS40K dataset (a) of SCENE-Net (b) against a CNN with similar architecture as baseline (c). Even though the CNN achieves a higher Recall on the majority of the samples, it classifies most vertical scene elements as towers, which ultimately leads to a poor segmentation of supporting towers. In contrast, SCENE-Net segments the body of towers and rejects other vertical objects that do not exhibit the prior knowledge encoded in the model. Thus, despite the CNN's greater expressive power and parameter count, SCENE-Net achieves superior segmentation performance, illustrating the strength of geometric inductive biases and the efficiency of our approach.

Comparison with state-of-the-art methods. To position SCENE-Net's performance within a broader context on TS40K, we also evaluate it against established 3D semantic segmentation architectures. In Table 2, we considered PointNet (Qi et al., 2017a), PointNet++ (Qi et al., 2017b), KPConv (Thomas et al., 2019), RandLA-Net (Hu et al., 2020), and the PointTransformer family (V1 through V3) (Zhao et al., 2021; Wu et al., 2022, 2024). We report Tower Intersection over Union (IoU), total number of trainable parameters, and define parameter efficiency as:

$$\text{Parameter Efficiency} = \frac{\text{Tower IoU}}{\log_{10}(\#\text{Parameters})}. \tag{13}$$

SCENE-Net achieves a parameter efficiency over an order of magnitude greater than any other method, while maintaining competitive Tower IoU. Although PointTransformer V3 attains the highest Tower IoU, it requires more than four million times the parameters of SCENE-Net. This efficiency is particularly important for resource-constrained applications, where model size can outweigh improvements in IoU.

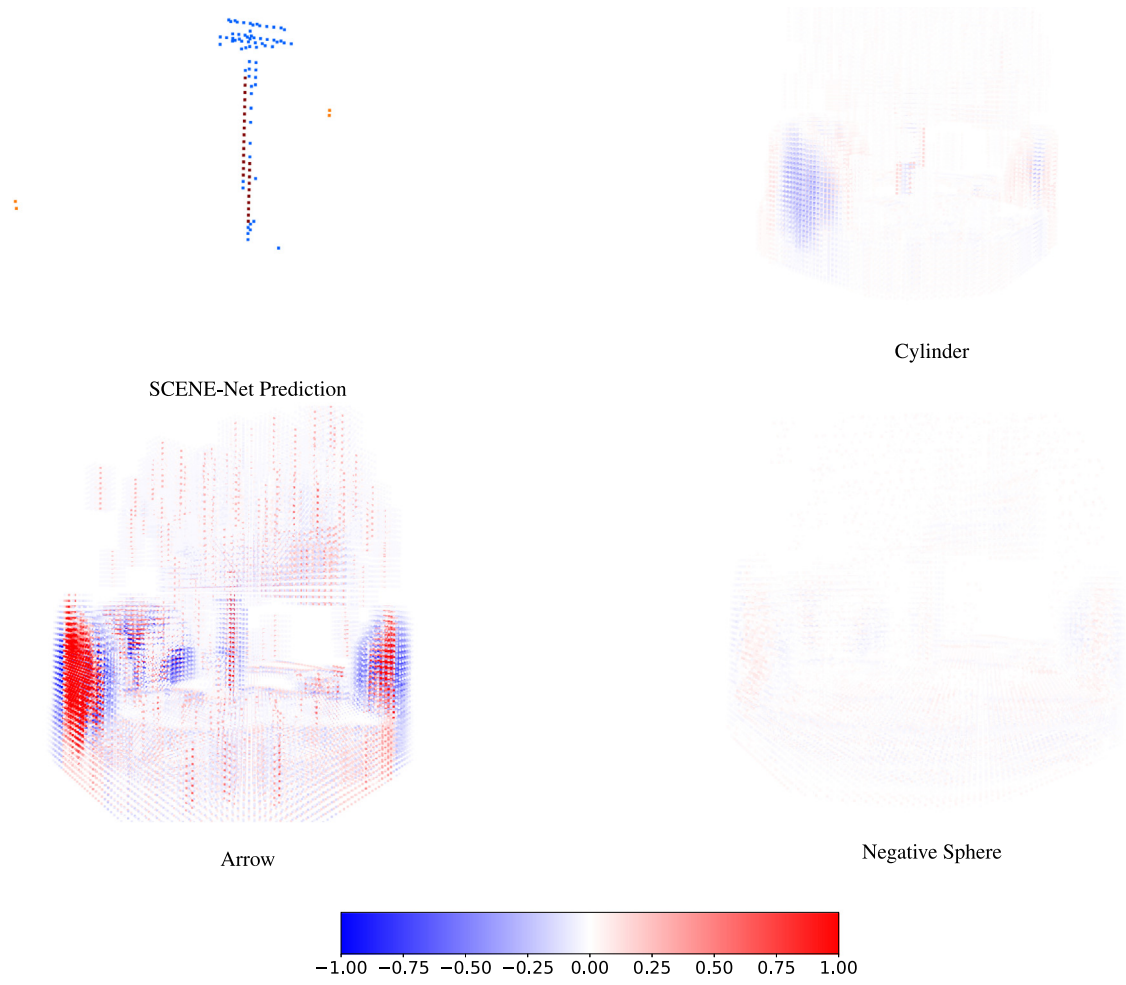

Fig. 5. *Post hoc* analysis of SCENE-Net. We can examine the activation of each geometric operator and correlate it to the detection of certain elements in the scene. We see that the Arrow is responsible for the most activation, while the Negative Sphere has a smaller absolute value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparison of Tower IoU and parameter efficiency on TS40K. Only models with both IoU and parameter counts reported are included for fair comparison. Methods are ordered by parameter efficiency.

Method	Tower IoU (%)	# Parameters (M)	Param. efficiency
SCENE-Net (ours)	58.0	**1.1e−5**	**55.68**
PTV2 (Wu et al., 2022)	61.7	12.8	8.67
PTV3 (Wu et al., 2024)	**65.1**	46.2	8.48
PTV1 (Zhao et al., 2021)	57.2	6.3	8.40
KPConv (Thomas et al., 2019)	42.7	14.9	5.95
PointNet++ (Qi et al., 2017b)	25.9	1.48	4.19
PointNet (Qi et al., 2017a)	8.6	0.40	1.86
RandLA-Net (Hu et al., 2020)	0.0	1.24	0.00

SCENE-Net on the SemanticKITTI. In Table 3, we present a detailed comparison of SCENE-Net's performance against leading models for 3D semantic segmentation on the SemanticKITTI benchmark along with qualitative results in Fig. 6. This comparison focuses on pole IoU, the number of parameters, and the ratio of pole IoU to the number of parameters. Input point clouds are voxelized with a voxel size of 0.05 m due to the sparsity and rectangular shape self-driving point clouds. Despite state-of-the-art models being trained in a multiclass setting while we focus only on pole-like structures, this comparison serves two purposes: (1) It demonstrates that SCENE-Net's specialization for pole detection is effective across various benchmarks and data characteristics. The SemanticKITTI dataset is a sparse, large-scale dataset with high occlusion and varying density, encompassing multiple 3D elements that SCENE-Net must ignore. (2) In terms of pole IoU, SCENE-Net achieves a remarkable performance of 57.5% with only 11 trainable parameters. The highest pole IoU recorded is 69.8% by TG-KD (Hai et al., 2025), with 2.78 million parameters. This result provides strong evidence that the GENEO-based observer generalizes beyond TS40K and adapts to other environments with significantly different data statistics and sensor properties. Although the GENEOs used in this work were specifically designed for TS40K, they remain effective in SemanticKITTI in pole detection without modification. This reflects the ability of our operators to transfer across domains as long as the target class retains consistent geometric traits. That said, further gains could be achieved by refining the GENEO set to better match the geometry of a new dataset. For example, SemanticKITTI features more fragmented pole instances due to higher occlusion. With SCENE-Net's white-box design, the observer's failures can be inspected and interpreted, allowing

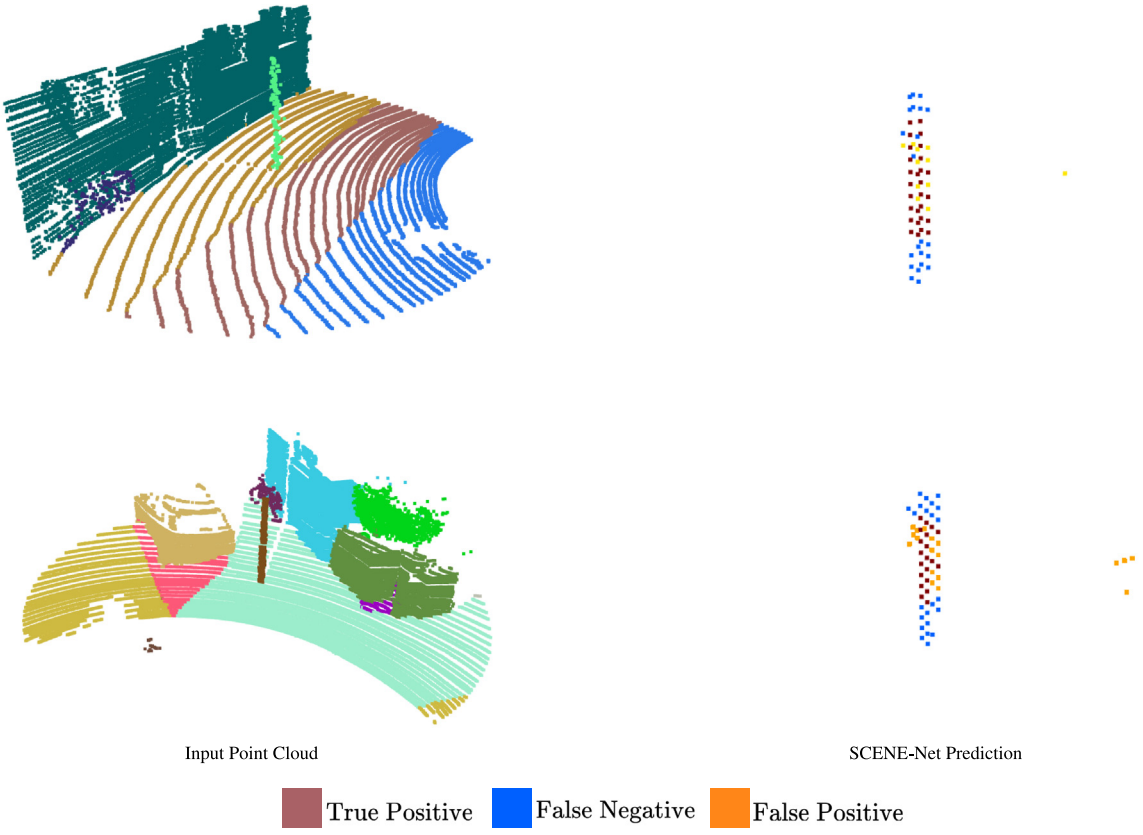

Fig. 6. Qualitative results of SCENE-Net on the SemanticKITTI dataset. The right column shows SCENE-Net's predictions, while the left column displays the input point clouds. The latter are colored according to the original semantic labels of the SemanticKITTI dataset to guarantee an easy distinction between 3D objects in the scene. SCENE-Net successfully identifies poles despite clutter, sparsity, and occlusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Semantic segmentation on *SemanticKITTI*: pole IoU, number of parameters, and parameter efficiency. Only methods for which both pole IoU and parameter count are available are included, to ensure fairness.

Method	Pole IoU (%)	#Parameters (M)	Param. efficiency
SCENE-Net (Ours)	57.5	**1.1e^{-5}**	**55.23**
TG-KD Student with KD (Hai et al., 2025)	**69.8**	2.78	10.83
TG-KD Student w/o KD (Hai et al., 2025)	63.8	2.78	9.90
PTV3 (Wu et al., 2024)	64.4	46.2	8.39
Cylinder3D (Zhu et al., 2021)	62.4	53.0	8.08
SalsaNext (Cortinhal et al., 2020)	54.3	6.7	7.95
JS3C-Net (Yan et al., 2021)	60.7	2.7	3.79
SparseConv (Graham et al., 2018)	57.9	2.7	3.61
SPVNAS (Tang et al., 2020)	64.3	12.5	3.46
RangeFormer (Kong et al., 2023)	66.4	24.3	3.09
RPVNet (Xu et al., 2021)	64.8	24.8	2.95
KPConv (Thomas et al., 2019)	56.4	14.9	2.92
RandLA-Net (Hu et al., 2020)	44.2	1.24	2.70
UniSeg (Liu et al., 2023a)	68.3	147.6	2.56
PointNet++ (Qi et al., 2017b)	6.0	1.48	0.84

new GENEOs to be crafted post hoc. Overall, leveraging the relationship between 3D points with transparent geometric priors could lead to more significant performance improvements in state-of-the-art methods than simply increasing model size. As shown in Table 3, increasing model size leads to diminishing returns.

SCENE-Net is robust to noisy inputs, labels, and occlusions. Robustness is an essential requirement in 3D scene understanding tasks, particularly with noisy labels and LiDAR artifacts present in real-world datasets such as TS40K. In this section, we assess SCENE-Net's resilience to three common sources of uncertainty: (1) geometric noise in point clouds, (2) partial input due to sparsity or occlusion, and (3) label noise in the ground truth annotations.

To simulate sensor noise and structural irregularities, we injected test-time additive noise into the spatial coordinates (x, y, z) of the input point clouds and assessed our previously trained model. We experimented with three noise distributions: Normal, Laplace, and Cauchy, each applied using three different standard deviations. As shown in Table 4, SCENE-Net retains high precision under low noise levels ($\sigma = 0.01$), and degrades as the noise intensity increases. Under high-variance or heavy-tailed perturbations (e.g., $\sigma = 0.1$ or Cauchy noise), tower geometry becomes unrecoverable and SCENE-Net abstains from

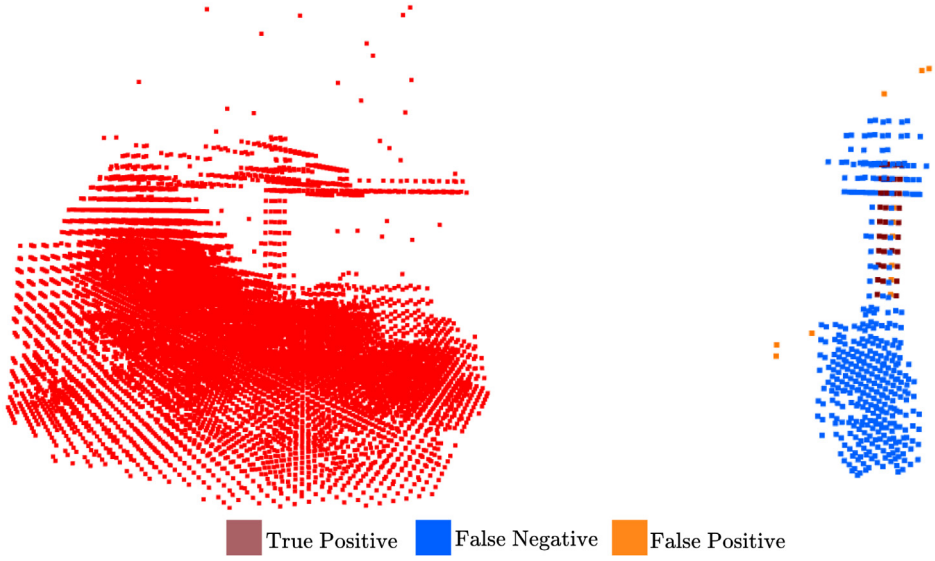

Fig. 7. SCENE-Net is robust to label noise. Left: ground truth includes a mislabeled road. Right: SCENE-Net prediction suppresses the ground patch and highlights only the tower.

Table 4
SCENE-Net performance under different test-time noise conditions. Coordinates are normalized to $[0, 1]$ prior to noise injection.

Noise type	Precision (%)	Recall (%)
No noise (base model)	82.0	13.0
Normal (0, 0.01)	81.1	11.6
Normal (0, 0.05)	50.4	2.7
Normal (0, 0.1)	0.0	0.0
Laplace (0, 0.01)	79.7	9.3
Laplace (0, 0.05)	47.2	1.0
Laplace (0, 0.1)	0.0	0.0
Cauchy (0, 0.01)	72.1	6.2
Cauchy (0, 0.05)	14.4	1.8
Cauchy (0, 0.1)	0.0	0.0

Table 5
SCENE-Net performance under random point dropout. Each dropout rate was applied at test time.

Dropout probability	Precision (%)	Recall (%)
0%	82.0	13.0
5%	81.0	13.0
10%	81.0	13.0
30%	80.8	12.5
50%	80.0	12.0

prediction entirely since the geometric patterns in the scene are inconsistent with known tower structures. We view this behavior as desirable because it reflects the conservative nature of the model, avoiding false positive predictions.

We also tested SCENE-Net's robustness to point sparsity by randomly dropping input points with dropout probabilities ranging from 0% to 50%. This simulates occlusions or sensor dropouts common in LiDAR data. As Table 5 shows, SCENE-Net remains robust even when 50% of the points are removed. This resilience stems from two properties of our design: (i) supporting towers comprise less than 1% of the total points in the TS40K dataset and, thus, are statistically unlikely to be entirely dropped; (ii) our binary voxelization treats any non-empty voxel as active, preserving tower structure despite point sparsity.

Lastly, beyond geometric degradation, annotation noise presents a significant challenge in supervised 3D learning. The TS40K dataset contains a high proportion of mislabeled voxels: approximately 50% of those labeled as "supporting tower" actually correspond to surrounding ground or road regions, rather than structural tower elements. As noted in Lavado et al. (2025), these noisy labels are due to annotation overflow: conservative labeling that extends beyond the target structure. This noise is not random but systematic, it stems from safety-driven inspection protocols rather than from a machine learning-oriented labeling standard. Consequently, several structures labeled as "tower" do not adhere to the expected pole-like geometry, introducing a mismatch between labels and the expected semantic group. Despite being trained on this imperfect supervision, SCENE-Net remains focused on the tower body and ignores irrelevant annotations. Fig. 7 shows a representative example: although the ground patch is labeled as tower, SCENE-Net activates only on the actual structure. This behavior explains the relatively low recall but very high precision reported in Table 1: SCENE-Net does not chase false positives even when they match the ground-truth label.

This robustness to both noisy input and noisy supervision arises from SCENE-Net's design: the GENEO observers embed meaningful geometric priors, which act as built-in regularizers. As a result, SCENE-Net does not overfit to noisy labels or ambiguous inputs and provides stable predictions in adverse conditions. We believe this is a major advantage of using interpretable, structured priors in 3D semantic segmentation.

Efficiency analysis. We evaluated the resource efficiency and learning process of SCENE-Net in comparison with representative baselines, using a single NVIDIA RTX 4070 GPU with 8 GB of RAM. All models were trained for 100 epochs with a batch size of 16. For SCENE-Net and a baseline CNN, input point clouds were voxelized into 64^3 grids (around 1.0 MB per sample, or 16 MB per batch), whereas other baselines operated directly on raw point clouds of 10k points (around 0.12 MB per sample, or 2 MB per batch). As summarized in Table 6, SCENE-Net demonstrates extreme parameter efficiency, requiring only 11 trainable weights, compared to over 2000 for the baseline CNN and several orders of magnitude more for PointNet++, KPConv, and Point Transformer v3. This dramatic reduction in parameters translates into lower GPU memory usage during training (peak 2.2 GB for SCENE-Net vs. 2.5 GB for the baseline CNN and up to 7.4 GB for the transformer model) and faster inference (3.1 ms per sample, versus 12–36 ms for state-of-the-art approaches). Despite a slightly longer per-epoch training time due to recomputation of GENEO kernels (2.3 min vs. 1.4 min for the baseline CNN), SCENE-Net converges much more

Table 6
Comparison of efficiency metrics between SCENE-Net and baselines. All models were trained on an NVIDIA RTX 4070 GPU with 8 GB RAM, for 100 epochs with a batch size of 16.

Method	Params (M)	Train time/epoch (m)	Time to converge (h)	Peak GPU mem (MB)	Inference latency (ms)
SCENE-Net (ours)	1.1×10^{-5}	2.3	**0.5**	**2128**	3.1
Baseline CNN	0.0022	**1.4**	1.1	2512	**2.9**
PointNet++	1.48	18.6	19.5	3201	12.5
KPConv	14.9	20.1	28.4	6589	20.3
Point Transformer v3	46.2	25.6	34.7	7454	35.7

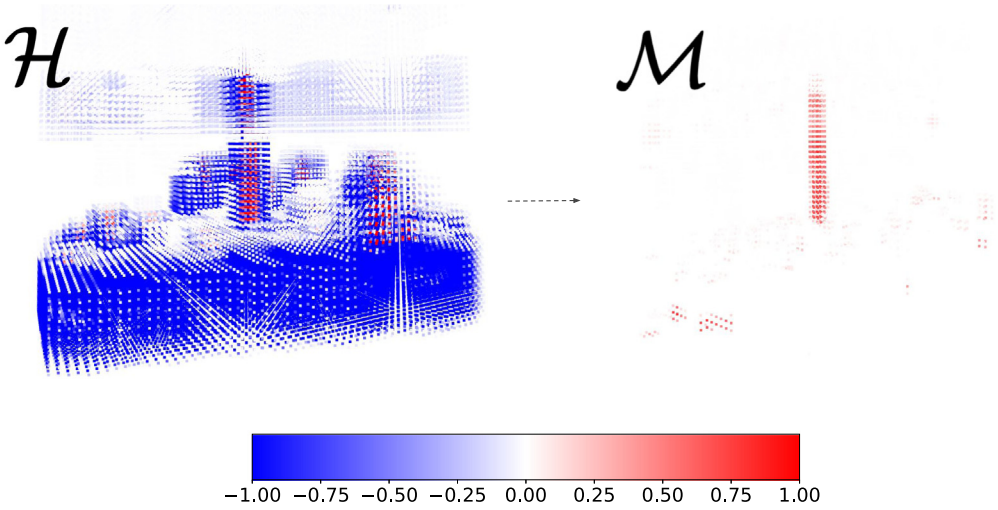

Fig. 8. Transformation of SCENE-Net's GENEO observer $\mathcal{H}$ into a tower probability model $\mathcal{M}$ as defined in Eq. (4). The left side shows the convex sum of three base GENEOs (cylinder, arrow, and negative sphere) into a single observer $\mathcal{H}$, which activates on pole-like structures while suppressing unrelated elements in the scene. We convert $\mathcal{H}$ into a normalized probability map $\mathcal{M}$ shown on the right. This discards negative responses and smoothly scales positive ones into $[0, 1]$, yielding a semantic density estimate for supporting towers.

quickly, reaching stable performance within 5 epochs, compared to roughly 20 for the baseline CNN and near the end of training for larger models. The combined efficiency in parameters, memory, and training time highlights SCENE-Net as highly suitable for low-resource environments and scarce-data regimes.

5.2.3. RQ3: GENEOs as inductive biases for 3D point clouds

SCENE-Net is a complex observer built from simple shapes. SCENE-Net is a white-box model composed of simple, interpretable geometric operators that collectively form a powerful semantic observer. Each base operator is designed to detect specific geometric patterns such as cylindrical or arrow-like shapes, and is constrained to be equivariant with respect to a transformation group that preserves verticality (e.g., rotations about the z-axis and translations in the xy-plane). In Fig. 8, we show how our three GENEO kernels are combined into a single observer $\mathcal{H}$, whose output highlights pole-like structures in the 3D scene while suppressing unrelated elements such as vegetation or terrain. Because $\mathcal{H}(x)$ is geometrically meaningful and directly signals the presence of poles, we transform its output into a tower probability map $\mathcal{M}(x)$ using the function $\mathrm{ReLU}(\tanh(\mathcal{H}(x)))$ as formalized in Eq. (4). This transformation eliminates all negative activations, corresponding to shapes that contradict GENEO priors, and scales positive responses into the range $[0, 1]$. The result is a meaningful, bounded density function indicating how likely each voxel is to contain a tower-like structure. Although the current implementation targets pole-like structures, the architecture is modular and easily extensible. New GENEOs can be added to target different geometries, such as planar structures or bent forms. For example, disk-shaped GENEOs could capture horizontal surfaces, and learnable rotation matrices could generalize angular configurations beyond verticality. These extensions would continue to preserve equivariance as long as the new GENEOs conform to the same transformation group, making then GENEOs valuable instruments for application-specific knowledge engineering.

Ablation study on GENEOs. In this section, we conduct an ablation study on SCENE-Net's architecture by varying the number of instances of each GENEO type: Cylinder, Arrow, and Negative Sphere. All models are evaluated on the TS40K validation set, and the corresponding GENEO configurations are listed in the second to fourth columns of Table 7. Models A and B use only a single GENEO type (Cylinder or Arrow), which results in poor segmentation performance, showing that no single geometric prior is sufficient on its own. Model B, which uses only the Negative Sphere GENEO, performs better than A by helping suppress vegetation, but still lacks the specificity needed for precise tower detection. Models C and D combine the Negative Sphere with either the Cylinder or Arrow, improving the model's ability to detect pole-like shapes while suppressing background clutter. However, the best performance is achieved by model E, the standard SCENE-Net configuration, which integrates all three GENEO types in a balanced combination. Finally, models F and G test whether increasing the number of GENEO instances improves results. Surprisingly, this leads to degraded performance, likely due to over-redundancy and dilution of convex weights across the kernels. This confirms that the original SCENE-Net configuration (model E) offers the best trade-off between expressivity and selectivity.

Voxel resolution and kernel size in SCENE-Net. One of the key challenges in voxel-based models is the computational burden of applying 3D convolutions over high-resolution voxel grids with large kernels. In traditional CNNs, both memory and training time grow cubically with kernel size and resolution. SCENE-Net avoids this limitation by defining its geometric observers as continuous operators. These GENEOs are independent of voxel discretization and are simply sampled onto the grid during training and inference. As a result, SCENE-Net's behavior does not depend on a specific resolution or kernel size.

To investigate this property, we take advantage of our trained SCENE-Net on voxel grids of size 64^3 and kernel size of $9 \times 5 \times 5$

Table 7
Ablation study of SCENE-Net on the TS40K validation set. Each row corresponds to a different configuration of the model, where the second to fourth columns indicate the number of Cylinder, Arrow, and Negative Sphere GENEOs used, respectively. We report segmentation performance (Precision, Recall, IoU) for each configuration to assess the contribution of different operator combinations.

Model	# Cylinder	# Arrow	# NegSphere	Precision (%)	Recall (%)	IoU (%)
A	1	0	0	0	0	0
B	0	1	0	0	0	0
C	1	0	1	34	1	12
D	0	1	1	13	1	8
E (Ours)	**1**	**1**	**1**	**82**	**13**	**58**
F	2	2	2	56	16	53
G	3	3	3	37	22	56

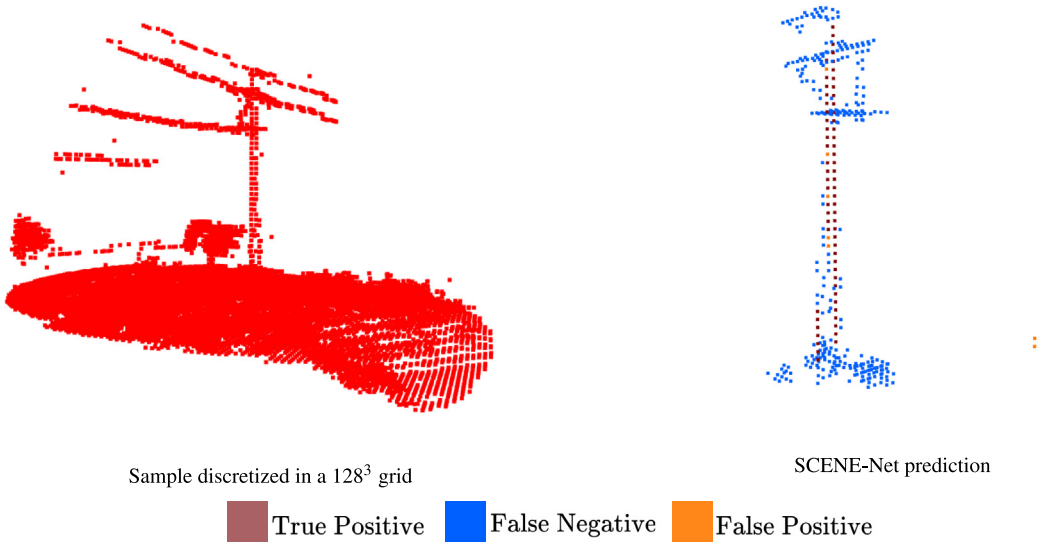

Fig. 9. SCENE-Net inference on a 128^3 grid using a kernel size of $12 \times 5 \times 5$, despite being trained at 64^3 with kernel size 9^3. This demonstrates resolution-agnostic behavior made possible by its continuous formulation.

Table 8
Effect of voxel grid resolution on SCENE-Net performance on the TS40K dataset (inference only, model trained at 64^3 with kernel size $(5, 5, 5)$).

Grid resolution	Precision (%)	Recall (%)
32^3	59.2	7.8
48^3	71.1	11.0
64^3	**79.7**	**11.6**
96^3	78.5	10.1
128^3	76.3	9.9

Table 9
Effect of GENEO kernel size on SCENE-Net performance on the TS40K dataset (inference only, model trained at 64^3 with kernel size $(5, 5, 5)$).

Kernel size (z,x,y)	Precision (%)	Recall (%)
(3, 3, 3)	59.1	4.2
(5, 5, 5)	79.7	11.6
(7, 7, 7)	80.4	11.0
(9, 9, 9)	**82.6**	9.3
(9, 5, 5)	82.5	**11.9**
(12, 5, 5)	77.4	8.2
(5, 9, 9)	63.6	6.5

and evaluated it at higher resolutions, such as 128^3, without retraining. As shown in Fig. 9, SCENE-Net retains its ability to localize tower structures in this finer discretization, using a larger kernel of size $12 \times 5 \times 5$. This qualitative result supports the hypothesis that SCENE-Net's behavior generalizes across resolutions due to its continuous formulation. To further quantify this, we conducted an ablation study where a model trained at 64^3 with kernel size $(5, 5, 5)$ was evaluated at several voxel resolutions. As shown in Table 8, performance peaks at the training resolution, with modest degradation at both coarser and finer grids. This behavior reflects the natural trade-off between spatial detail and distribution shift across grid scales, and confirms that SCENE-Net generalizes robustly across input resolutions.

In a complementary ablation, we tested the same trained model on varying kernel sizes. Results in Table 9 show that different GENEO shapes lead to different trade-offs in precision and recall, with the best configuration being a kernel of shape $(9, 5, 5)$. Notably, this tuning happens entirely at inference time, without retraining. The ability to decouple kernel design from training is unique to SCENE-Net and reinforces its utility as a white-box approach.

6. Conclusions

In this paper, we presented SCENE-Net, an interpretable and computationally efficient approach for segmenting 3D point clouds. By leveraging Group Equivariant Non-Expansive Operators (GENEOs), SCENE-Net incorporates fundamental geometric priors as inductive biases, diverging from the commonly employed black-box models in the domain. With only a few significant parameters, SCENE-Net shows that high performance and interpretability can coexist, impacting fields like infrastructure maintenance. SCENE-Net's effectiveness highlights the value of integrating geometric biases in deep learning. Although it requires initial manual definition of GENEOs, it is possible to define a library of basic geometric priors that can potentially adapt to multiple

tasks. An exciting direction for future work is extending SCENE-Net to multi-class segmentation. This could be achieved by composing multiple GENEO observers, each tuned to different geometric patterns, and combining their outputs through a lightweight classifier, possibly considering also the presence of major discontinuities in the patterns (Micheletti et al., 2020). Such an approach would preserve SCENE-Net's interpretability while enabling richer semantic understanding across object categories. This also opens the door to the use of GENEOs as rich feature extraction tools for state-of-the-art methods to improve performance.

CRediT authorship contribution statement

Diogo Lavado: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alessandra Micheletti:** Writing – review & editing, Methodology, Formal analysis. **Giovanni Bocchi:** Writing – review & editing, Formal analysis. **Patrizio Frosini:** Writing – review & editing, Formal analysis. **Cláudia Soares:** Writing – review & editing, Methodology, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by DM 351-2022 of the Italian Ministry of University and Research and with partial support from the Italian INDAM – GNAMPA group, by FCT I.P., through the strategic project NOVA LINCS (UIDB/04516/2020), "Sustainable Stone by Portugal" N° C644943391-00000051/40 co-funded by Recovery and Resilience Plan and NextGeneration EU, and TaRDIS Horizon Europe (101093006).

Appendix A. Proofs of non-expansiveness

A.1. Proof of non-expansiveness of the Arrow GENEO

Take the definition of the Arrow GENEO as:

$$\begin{aligned}&\int_0^h \int_{\mathbb{R}^2} g_{Ar}(x)dx \\ &= h_c \int_{\mathbb{R}^2} e^{-\frac{1}{2\sigma^2}(\|\tilde{x}-\tilde{c}\|^2-r^2)^2} d\tilde{x} \; + \\ &\quad (h-h_c)\int_{\mathbb{R}^2} e^{-\frac{1}{2\sigma^2}(\|\tilde{x}-\tilde{c}\|^2-r_h(x))^2} d\tilde{x}\end{aligned} \tag{A.1}$$

where $\tilde{x}$ and $\tilde{c}$ are the projections on the first two coordinates, and $r_h(x) = (h-\pi_3(x))r_c \tan(\beta\pi)$ defines a specific radius for a 3D point at a certain height. In 4.2.1, we have proven that the first integral

$$h_c \int_{\mathbb{R}^2} e^{-\frac{1}{2\sigma^2}(\|\tilde{x}-\tilde{c}\|^2-r^2)^2} d\tilde{x}$$

is bounded by a constant $C = h_c(A_B C_0 + C_1)$, where h_c is the height of the cylinder, and $A_B C_0$ and C_1 define bounds when integrating inside and outside the disk $B(0,t^\star)$.

Since, $r_h(x) : \mathbb{R}^3 \to \mathbb{R}$ produces a radius for a given 3D point, the cone defined by the second integral can be encased in a cylinder with radius

$$r_{\max} = \max_{x\in\mathbb{R}^3} (h-\pi_3(x))r_c \tan(\beta\pi).$$

Thus, we can follow the same rationale as in 4.2.1 to prove that the second integral is also bounded. Specifically

$$(h-h_c)\int_{\mathbb{R}^2} e^{-\frac{1}{2\sigma^2}(\|\tilde{x}-\tilde{c}\|^2-r_{\max})^2} d\tilde{x} \le (h-h_c)(A_B C_2 + C_3), \tag{A.2}$$

where

$$C_2 = \max_{B(0,t^\star)} e^{-\frac{1}{2\sigma^2}(\|z\|^2-r_{\max}^2)^2}.$$

and

$$C_3 = \int_{\bar{B}(0,t^\star)} e^{-\frac{\|y\|^2}{2\sigma^2}} dy$$

This entails that the overall integral is bounded by $C_{Ar} = C + (h-h_c)(A_B C_2 + C_3)$. The constant C_{Ar} can be used to normalize $g_{Ar}(x)$, making it non-expansive.

A.2. Proof of non-expansiveness of the negative sphere GENEO

Take the definition of the Negative Sphere GENEO as:

$$\begin{aligned}&\int_{\mathbb{R}^3} g_{NS}(x)dx \\ &= \int_{\mathbb{R}^3} -\omega\, e^{\frac{-1}{2\sigma^2}(\|\tilde{x}-\tilde{c}\|^2-r^2)^2}\end{aligned} \tag{A.3}$$

where $\tilde{x}$ and $\tilde{c}$ are the projections on the first two coordinates and $\omega \in (0,1]$ is a negative factor. Since we are proving Non-expansiveness over the L_1 norm, specifically

$$\int_{\mathbb{R}^3} |g_{NS}(x)|dx < \infty, \tag{A.4}$$

the negative sign of the Negative Sphere can be disregarded. We start by a variable transformation $z = \tilde{x} - \tilde{c}$ yielding, as the inner integral

$$X = \int_{\mathbb{R}^3} \omega e^{-\frac{1}{2\sigma^2}(\|\tilde{x}-\tilde{c}\|^2-r^2)^2} dx = \int_{\mathbb{R}^3} \omega e^{-\frac{1}{2\sigma^2}(\|z\|^2-r^2)^2} dz. \tag{A.5}$$

We now consider the fact that there is a positive scalar $t^\star > 0$, such that for all $t > t^\star$, $t^2 - r^2 > t$. Thus, we break the integral in (A.5) in two terms, one integrating over the ball centered in zero with radius $t^\star$, $B(0,t^\star)$, and another integrating over the complementary region, outside the ball, $\bar{B}(0,t^\star)$

$$X = \int_{B(0,t^\star)} -\omega e^{-\frac{1}{2\sigma^2}(\|z\|^2-r^2)^2} dz + \int_{\bar{B}(0,t^\star)} \omega e^{-\frac{1}{2\sigma^2}(\|z\|^2-r^2)^2} dz.$$

The first integral

$$X_0 = \int_{B(0,t^\star)} \omega e^{-\frac{1}{2\sigma^2}(\|z\|^2-r^2)^2} dz$$

exists and is bounded by the volume of the ball V_B multiplied by the maximum C_0 defined as

$$C_0 = \max_{B(0,t^\star)} \omega e^{-\frac{1}{2\sigma^2}(\|z\|^2-r^2)^2}.$$

Here, by the Weierstrass extreme value theorem, C_0 is a finite number because the function to be maximized is a continuous function over a compact set. The second integral can also be bounded. According to the definition of $t^\star$,

$$\begin{aligned}\|y\| > t^\star &\implies \|y\|^2 - r^2 \ge \|y\| && \text{(A.6)}\\ &\implies (\|y\|^2 - r^2)^2 \ge \|y\|^2 && \text{(A.7)}\\ &\implies -(\|y\|^2 - r^2)^2 \le -\|y\|^2 && \text{(A.8)}\\ &\implies e^{-\frac{(\|y\|^2-r^2)^2}{2\sigma^2}} \le e^{-\frac{\|y\|^2}{2\sigma^2}} && \text{(A.9)}\\ &\implies \omega e^{-\frac{(\|y\|^2-r^2)^2}{2\sigma^2}} \le \omega e^{-\frac{\|y\|^2}{2\sigma^2}}. && \text{(A.10)}\end{aligned}$$

the first inequality entails the second because $\|y\|$ is a non-negative number. The integral

$$C_1 = \int_{\bar{B}(0,t^\star)} \omega e^{-\frac{\|y\|^2}{2\sigma^2}} dy$$

can be easily computed. The overall integral is bounded by $C =$

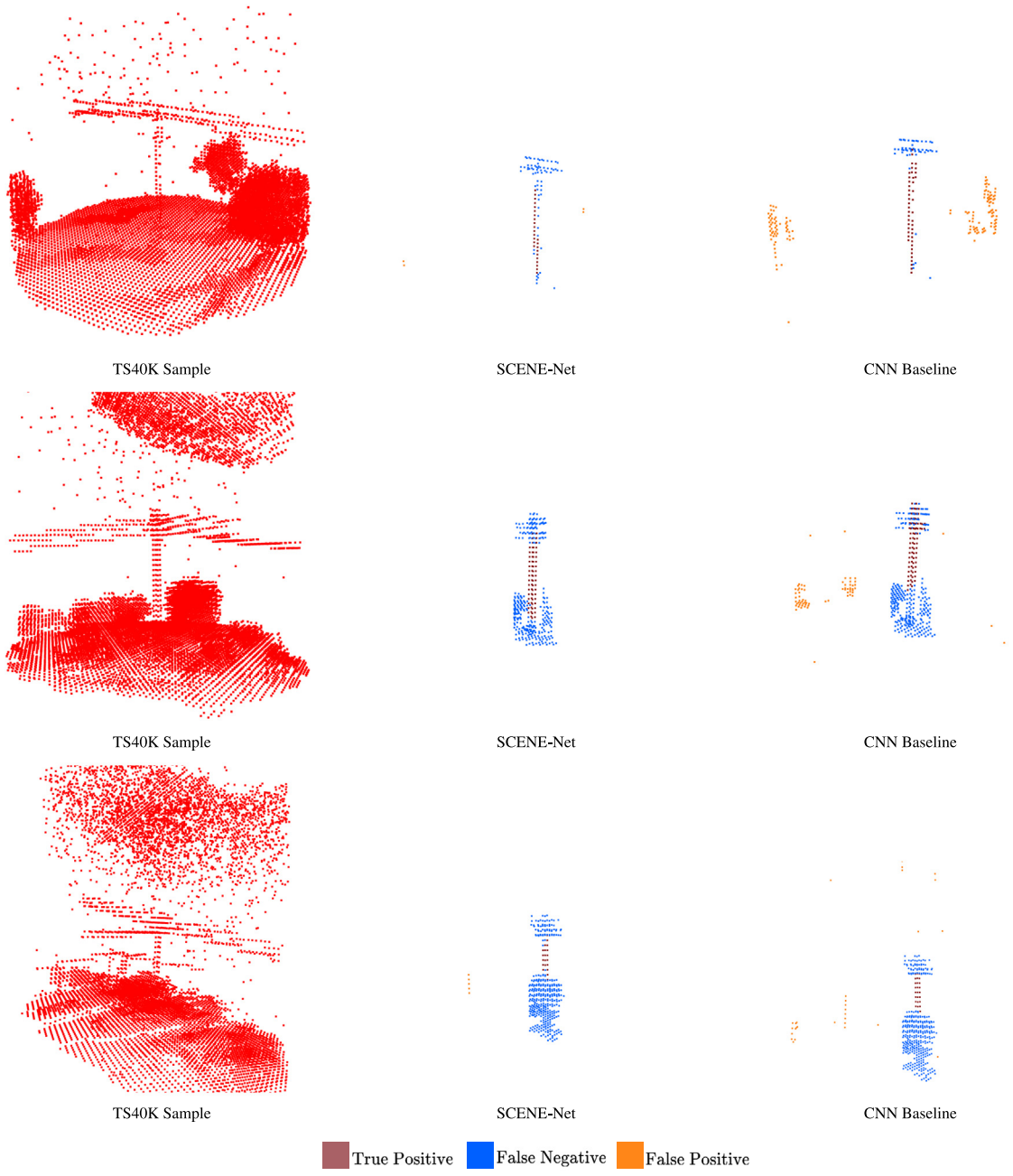

Fig. D.10. Qualitative results of SCENE-Net on the testing set of TS40K, against a CNN with a similar architecture. Note that, in the last example, SCENE-Net clearly identifies a second unlabeled tower, whereas CNN identifies both the second tower and vegetation as towers.

$V_B C_0 + C_1$. The constant C can be used to normalize $g_{NS}(x)$, making it non-expansive.

Appendix B. Power line segmentation from 3D point clouds

Power line inspection is generally performed by on-site maintenance personnel and manned helicopters that examine the power grid with portable devices or the naked eye. These methods are costly, inefficient, and demanding for staff. Thus, process automation is crucial for operators. To this end, unmanned aerial vehicles (UAVs) carrying LiDAR sensors are deployed to scan the power grid and capture a 3D point cloud representation of the environment. In the work (Ding et al., 2021), the authors combine SLAM algorithms with multi-sensor data to patrol the electrical grid with UAVs. This method employs a multi-view-based approach to point cloud segmentation, so the 3D reconstructions from 2D raster maps usually introduce information loss and decrease in accuracy. Alternative methods such as Guo et al. (2019) project point clouds onto the xy-plane in order to cluster them to segment power lines. This approach does not consider ground and irregular terrain and focuses on segmenting incomplete power lines. Other methods take advantage of fine-grained elevation statistics of the original point cloud and xy-plane projections like (Tao et al., 2019). The proposals above focus on the segmentation of high-voltage power lines and disregard their supporting towers. By incorporating prior knowledge, our proposal segments supporting towers of any voltage. Not only are these structures also subject to inspections, but they serve as a point of reference for the location of power lines. In addition, by taking into account raw 3D scenes, other scene elements, such as vegetation, can be segmented to assess the risk of contact between the power grid and the environment.

Appendix C. Cylinder detection in 3D point clouds

Several classical and recent methods have been proposed for detecting cylindrical structures in 3D point clouds, particularly in industrial or pipeline environments (Liu et al., 2013; Tran et al., 2015; Araújo and Oliveira, 2020; Lu et al., 2021). These approaches often rely on geometric fitting pipelines using RANSAC, curvature estimation, or projection-based simplifications. For example, Liu et al. (2013) reduce 3D cylinder detection to 2D circle estimation for structured environments, while Lu et al. (2021) decompose scenes via axis-aligned slicing to track cylinder candidates. Tran et al. (2015) and Araújo and Oliveira (2020) apply multi-stage refinement and connectivity-based heuristics to extract cylinders in dense, well-organized scans. While effective in controlled settings, these methods are limited when applied to noisy and cluttered outdoor data such as power grid inspection point clouds. They typically assume dense and consistent geometry, rely on hand-tuned parameters, or lack semantic awareness. In contrast, SCENE-Net encodes geometric priors using interpretable GENEOs and learns to identify semantically relevant pole-like structures while ignoring visually similar but irrelevant elements. Unlike geometric fitting methods tailored to isolated primitives, SCENE-Net offers a modular and extensible framework that can generalize across environments and adapt to new 3D structures by incorporating additional GENEOs without redesigning the full pipeline.

Appendix D. Additional qualitative results

In Fig. D.10, we can analyze the qualitative results on the TS40K dataset of SCENE-Net against a CNN is similar architecture as baseline. Even though the CNN achieves a higher Recall on the majority of the samples, it classifies most vertical scene elements as towers, which ultimately leads to a poor segmentation of supporting towers. In contrast, SCENE-Net segments the body of towers and rejects other vertical objects that do not exhibit the prior knowledge encoded in the model.

Data availability

Data will be made available on request.

References

Araújo, A.M., Oliveira, M.M., 2020. Connectivity-based cylinder detection in unorganized point clouds. Pattern Recognit. 100, 107161.

Baxter, J., 2000. A model of inductive bias learning. J. Artificial Intelligence Res. 12, 149–198.

Bergomi, M.G., Frosini, P., Giorgi, D., Quercioli, N., 2019. Towards a topological–geometrical theory of group equivariant non-expansive operators for data analysis and machine learning. Nat. Mach. Intell. 1 (9), 423–433.

Bocchi, G., Ferri, M., Frosini, P., 2025a. A novel approach to graph distinction through GENEOs and permutants. Sci. Rep. 15, 6259.

Bocchi, G., Frosini, P., Micheletti, A., Pedretti, A., Gratteri, C., Lunghini, F., Beccari, A.R., Talarico, C., 2022. GENEOnet: A new machine learning paradigm based on group equivariant non-expansive operators. An application to protein pocket detection. arXiv preprint arXiv:2202.00451.

Bocchi, G., Frosini, P., Micheletti, A., Pedretti, A., Gratteri, C., Lunghini, F., Beccari, A.R., Talarico, C., 2025b. GENEOnet: statistical analysis supporting explainability and trustworthiness. Statistics 59 (4), 1037–1062.

Botteghi, S., Brasini, M., Frosini, P., Quercioli, N., 2020. On the finite representation of group equivariant operators via permutant measures. arXiv preprint arXiv:2008.06340.

Cascarano, P., Frosini, P., Quercioli, N., Saki, A., 2021. On the geometric and Riemannian structure of the spaces of group equivariant non-expansive operators. arXiv preprint arXiv:2103.02543.

Chefer, H., Gur, S., Wolf, L., 2021. Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 782–791.

Chen, Z., Bei, Y., Rudin, C., 2020. Concept whitening for interpretable image recognition. Nat. Mach. Intell. 2 (12), 772–782.

Cohen, T.S., Geiger, M., Weiler, M., 2019. A general theory of equivariant cnns on homogeneous spaces. Adv. Neural Inf. Process. Syst. 32.

Cohen, T., Welling, M., 2016. Group equivariant convolutional networks. In: International Conference on Machine Learning. PMLR, pp. 2990–2999.

Conti, F., Frosini, P., Quercioli, N., 2022. On the construction of group equivariant non-expansive operators via permutants and symmetric functions. Front. Artif. Intell. 5, 16.

Cortinhal, T., Tzelepis, G., Erdal Aksoy, E., 2020. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In: International Symposium on Visual Computing. Springer, pp. 207–222.

Ding, L., Wang, J., Wu, Y., 2021. Electric power line patrol operation based on vision and laser slam fusion perception. In: 2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering. AUTEEE, IEEE, pp. 125–129.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Fong, R.C., Vedaldi, A., 2017. Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3429–3437.

Goyal, A., Bengio, Y., 2022. Inductive biases for deep learning of higher-level cognition. Proc. R. Soc. A 478 (2266), 20210068.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3D semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9224–9232.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51 (5), 1–42.

Guo, T., Xu, L., Chen, Y., Liu, Y., Zhao, J., Luo, X., Wu, S., 2019. Research on point cloud power line segmentation and fitting algorithm. In: 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference. IAEAC, Vol. 1, IEEE, pp. 2404–2409.

Hai, L.T., Le, T.D., Ding, Z., Tian, Q., Hy, T.-S., 2025. Topology-guided knowledge distillation for efficient point cloud processing. arXiv preprint arXiv:2505.08101.

Hashemi, N.S., Aghdam, R.B., Ghiasi, A.S.B., Fatemi, P., 2016. Template matching advances and applications in image analysis. arXiv preprint arXiv:1610.07231.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11108–11117.

Hua, B.-S., Tran, M.-K., Yeung, S.-K., 2018. Pointwise convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 984–993.

Kong, L., Liu, Y., Chen, R., Ma, Y., Zhu, X., Li, Y., Hou, Y., Qiao, Y., Liu, Z., 2023. Rethinking range view representation for lidar segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 228–240.

Lai, X., Chen, Y., Lu, F., Liu, J., Jia, J., 2023. Spherical transformer for lidar-based 3d recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17545–17555.

Lavado, D., Santos, R., Coelho, A., Santos, J., Micheletti, A., Soares, C., 2025. Learning under noisy labels, spurious points, and diverse structures: TS40k, a 3D point cloud dataset of rural terrain and electrical transmission systems. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 7326–7336.

Lavie, I., Gur-Ari, G., Ringel, Z., 2024. Towards understanding inductive bias in transformers: A view from infinity. arXiv preprint arXiv:2402.05173.

Le, T., Duan, Y., 2018. Pointgrid: A deep network for 3D shape understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9204–9214.

Levi, M.Y., Gilboa, G., 2024. Fast and simple explainability for point cloud networks. arXiv preprint arXiv:2403.07706.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B., 2018. PointCNN: Convolution on x-transformed points. Adv. Neural Inf. Process. Syst. 31.

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.. Queue 16 (3), 31–57.

Liu, Y., Chen, R., Li, X., Kong, L., Yang, Y., Xia, Z., Bai, Y., Zhu, X., Ma, Y., Li, Y., et al., 2023a. Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21662–21673.

Liu, Z., Yang, X., Tang, H., Yang, S., Han, S., 2023b. Flatformer: Flattened window attention for efficient point cloud transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1200–1211.

Liu, Y.-J., Zhang, J.-B., Hou, J.-C., Ren, J.-C., Tang, W.-Q., 2013. Cylinder detection in large-scale point cloud of pipeline plant. IEEE Trans. Vis. Comput. Graphics 19 (10), 1700–1707.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Lu, Z., Mao, W., Dai, Y., Li, W., Su, Z., 2021. Slicing-tracking-detection: Simultaneous multi-cylinder detection from large-scale and complex point clouds. IEEE Trans. Vis. Comput. Graphics 28 (12), 4172–4185.

Matrone, F., Paolanti, M., Felicetti, A., Martini, M., Pierdicca, R., 2022. Bubblex: An explainable deep learning framework for point-cloud classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 15, 6571–6587.

Micheletti, A., Aletti, G., Ferrandi, G., Bertoni, D., Cavicchioli, D., Pretolani, R., 2020. A weighted χ^2 test to detect the presence of a major change point in non-stationary Markov chains. Stat. Methods Appl. 29 (4), 899–912.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. Pointnet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst. 30.

Rethage, D., Wald, J., Sturm, J., Navab, N., Tombari, F., 2018. Fully-convolutional point networks for large-scale point clouds. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 596–611.

Ribeiro, M., Leite, J., 2021. Aligning artificial neural networks and ontologies towards explainable ai. In: Proceedings of the AAAI Conference on Artificial Intelligence. 35, (6), pp. 4932–4940.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

Ribeiro, M.T., Singh, S., Guestrin, C., 2018. Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence. 32.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1 (5), 206–215.

Steininger, M., Kobs, K., Davidson, P., Krause, A., Hotho, A., 2021. Density-based weighting for imbalanced regression. Mach. Learn. 110 (8), 2187–2211.

Su, H., Jampani, V., Sun, D., Maji, S., Kalogerakis, E., Yang, M.-H., Kautz, J., 2018. Splatnet: Sparse lattice networks for point cloud processing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2530–2539.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: International Conference on Machine Learning. PMLR, pp. 3319–3328.

Tan, H., Kotthaus, H., 2022. Surrogate model-based explainability methods for point cloud nns. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2239–2248.

Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S., 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In: European Conference on Computer Vision. Springer, pp. 685–702.

Tao, G., Lianggang, X., Heng, Y., Liugui, Y., Jian, Z., Shaohua, W., Di, W., 2019. Study on segmentation algorithm with missing point cloud in power line. In: 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference. IMCEC, IEEE, pp. 1895–1899.

Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S., 2017. Segcloud: Semantic segmentation of 3D point clouds. In: 2017 International Conference on 3D Vision (3DV). IEEE, pp. 537–547.

Thomas, H., Qi, C.R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L.J., 2019. Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6411–6420.

Tran, T.-T., Cao, V.-T., Laurendeau, D., 2015. Extraction of cylinders and estimation of their parameters from point clouds. Comput. Graph..

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I., 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. arXiv preprint arXiv:1905.09418.

Wang, P.-S., 2023. Octformer: Octree-based transformers for 3d point clouds. ACM Trans. Graph. 42 (4), 1–11.

Wang, P.-S., Liu, Y., Guo, Y.-X., Sun, C.-Y., Tong, X., 2017. O-CNN: Octree-based convolutional neural networks for 3D shape analysis. ACM Trans. Graph. 36 (4), 1–11.

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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4840–4851.

Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H., 2022. Point transformer V2: Grouped vector attention and partition-based pooling. In: NeurIPS.

Wu, W., Qi, Z., Fuxin, L., 2019. Pointconv: Deep convolutional networks on 3D point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9621–9630.

Xu, J., Zhang, R., Dou, J., Zhu, Y., Sun, J., Pu, S., 2021. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16024–16033.

Yan, X., Gao, J., Li, J., Zhang, R., Li, Z., Huang, R., Cui, S., 2021. Sparse single sweep LiDAR point cloud segmentation via learning contextual shape priors from scene completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. 35, (4), pp. 3101–3109.

Zhang, Q., Wu, Y.N., Zhu, S.-C., 2018. Interpretable convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR.

Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V., 2021. Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16259–16268.

Zheng, T., Chen, C., Yuan, J., Li, B., Ren, K., 2019. Pointcloud saliency maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1598–1606.

Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D., 2021. Cylindrical and asymmetrical 3D convolution networks for lidar segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9939–9948.