SCENE-Net V2: Interpretable Multiclass 3D Scene Understanding with Geometric Priors

Diogo Lavado [1 2] **Cláudia Soares** [1] **Alessandra Micheletti** [2]

Editors: *S. Vadgama, E.J. Bekkers, A. Pouplin, S.O. Kaba, H. Lawrence, R. Walters, T. Emerson, H. Kvinge, J.M. Tomczak, S. Jegelka*

Abstract

In this paper, we present SCENE-Net V2, a new resource-efficient, **gray-box model** for multiclass 3D scene understanding. SCENE-Net V2 leverages Group Equivariant Non-Expansive Operators (GENEOs) to incorporate fundamental geometric priors as inductive biases, offering a more transparent alternative to the prevalent black-box models in the domain. This model addresses the limitations of its white-box predecessor, SCENE-Net, by expanding its applicability from pole-like structures to a wider range of datasets with detailed 3D elements. Our model achieves the sweet-spot between application and transparency: SCENE-Net V2 is a general method for object identification with interpretability guarantees. Our experimental results demonstrate that SCENE-Net V2 achieves competitive performance with a significantly lower parameter count. Furthermore, we propose the use of GENEO-based architectures as a feature extraction tool for black-box models, enabling an increase in performance by adding a minimal number of meaningful parameters. Our code is available in: https://github.com/dlavado/SCENE-Net-V2

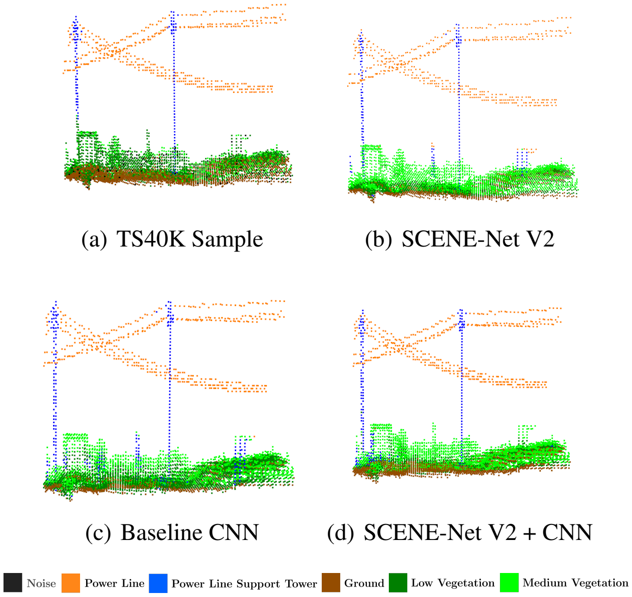

(a) TS40K Sample
(b) SCENE-Net V2
(c) Baseline CNN
(d) SCENE-Net V2 + CNN

Figure 1. 3D Semantic Segmentation of the TS40K Dataset: For the sample in (a), SCENE-Net V2 accurately detects the tower's body in (b), while a similar CNN model misclassifies medium vegetation as part of the tower in (c). In (d), using SCENE-Net V2 for geometric feature extraction combined with the same CNN architecture significantly improves segmentation performance. This is achieved by adding 540 meaningful parameters to the CNN.

1. Introduction

Recent advancements in 3D scene understanding have concentrated on enlarging models and datasets to enhance performance. For instance, Point Transformer V2 (Wu et al., 2023a) has more than tripled its parameter count when compared to its predecessor, Point Transformer V1 (Wu et al., 2022). In addition, benchmarks such as (AF)2-S3Net (Cheng et al., 2021) and 2DPASS (Yan et al., 2022) explore channel fusion (i.e., 2D projections, raw 3D point clouds and 3D voxel grids) to increase performance, which entails a severe toll in computational requirements and memory footprint. While this approach has yielded notable benefits, it often does not take into consideration the critical geometric information intrinsic to 3D point clouds. Harnessing this geometric data is essential for accurately interpreting and understanding three-dimensional environments and can be the key to drive innovation in crucial applications, such as

[1] Department of Environmental Science, University of Milan, Milan, Italy [2] Department of Computer Science, NOVA School of Science and Technology, Lisbon, Portugal. Correspondence to: Diogo Lavado <d.lavado@campus.fct.unl.pt>.

Proceedings of the Geometry-grounded Representation Learning and Generative Modeling at the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR Vol Number, 2024. Copyright 2024 by the author(s).

autonomous driving and environmental monitoring. Furthermore, these applications highlight the necessity for models that are easy to implement, efficient with data, and transparent, to guarantee their ethical and responsible use (Lipton, 2018; Guidotti et al., 2018; Doshi-Velez & Kim, 2017).

In this work, we present an interpretable 3D semantic segmentation model that leverages geometric priors within a deep neural network. This model not only provides transparency but also outperforms comparable black-box models in terms of accuracy and efficiency. To achieve such geometric priors, we leverage Group Equivariant Non-Expansive Operators (GENEOs) (Bergomi et al., 2019; Cascarano et al., 2021). GENEOs serve as building-blocks to describe Machine Learning agents as sets of operators acting on some input data. These operators provide a measure of the world, analogous to patterns learnt by CNN kernels. However, GENEOs are not blind to the underlying geometry of 3D scenes. They are parameterized with meaningful geometric features, such as the radius of a cylinder or the focal points of an ellipsoid, that define signature shapes found in 3D environments. SCENE-Net (Lavado et al., 2023) introduced the first application of GENEOs to 3D scene understanding. By combining three GENEOs into a learning agent with only 11 trainable parameters, SCENE-Net is able to detect pole-like structures in different datasets with great efficiency in training. Throughout this study, we evaluate the limitations of SCENE-Net (Lavado et al., 2023) and introduce a new, more powerful, and interpretable GENEO-based model named **SCENE-Net V2**. Our model enhances SCENE-Net by incorporating novel GENEO kernels, a more sophisticated design, and an architecture capable of performing 3D semantic segmentation across multiple classes, not just poles.

Our **contributions** include: **(1)** Proposing SCENE-Net V2, the first gray-box model for multiclass 3D semantic segmentation. **(2)** Introducing novel GENEO kernels with general geometric priors that aid the detection of various 3D elements. **(3)** Studying the use of SCENE-Net V2 as a geometric feature extraction tool for black-box models.

2. Related Work

2.1. Point Cloud Semantic Segmentation

Semantic segmentation at the scene level aims to divide a 3D point cloud into subsets based on the semantic meanings assigned to individual points. This process requires a comprehensive understanding that simultaneously considers the overarching geometric structure and the intricate details of each point. Semantic segmentation methodologies are typically categorized into three paradigms (Guo et al., 2020): **(1) Projection-based methods** (Su et al., 2015; Lawin et al., 2017; Yang & Wang, 2019; Lyu et al., 2020): These approaches utilize established 2D CNN frameworks to infer 3D semantics. However, the projection of point clouds onto 2D images can result in the loss of essential geometric information. **(2) Discretization-based methods** (Choy et al., 2019; Zhou & Tuzel, 2018; Le & Duan, 2018; Meng et al., 2019; Zhang et al., 2020): These models employ 3D CNN architectures. While effective, they often face scalability issues due to their significant computational and memory demands. **(3) Point-based methods** (Qi et al., 2017a;b; Li et al., 2018; Thomas et al., 2019; Hu et al., 2020; Kong et al., 2023; Lai et al., 2023; Wu et al., 2022; 2023a;b): These techniques use shared MLPs and transformers to learn semantics at the individual point level. Unlike voxel and projection-based methods, point-based architectures retain the semantics of each 3D point and achieve state-of-the-art performance across major datasets. Discretization-based methods are often computationally intensive, especially when handling high-resolution 3D voxel grids, whereas point-based approaches require complex neighbor-searching algorithms to extract local information. To address these challenges, we propose a voxel-based architecture that offers a time-efficient solution for high-resolution grids, supporting sizes from 64^3 to 256^3.

2.2. Explainable Machine Learning

Explainability in machine learning is vital, particularly for high-stakes applications where understanding model decisions is critical (Lipton, 2018; Guidotti et al., 2018; Doshi-Velez & Kim, 2017). Approaches to explainability are generally divided into *post hoc* explainability and intrinsic interpretability. *Post hoc* techniques, such as LIME (Ribeiro et al., 2016), meaningful perturbations (Fong & Vedaldi, 2017), anchors (Ribeiro et al., 2018), and ontologies (Ribeiro & Leite, 2021; Barbiero et al., 2022), explain black-box models' predictions by linking inputs to outputs. While flexible and model-agnostic, these methods often fail to provide deep causal insights and can misinterpret feature significance, sometimes equating irrelevant inputs with meaningful ones (Rudin, 2019). In contrast, intrinsic interpretability involves embedding explanations within the model's architecture, as seen in decision trees and linear models. These white-box models typically prioritize transparency over complexity, simplifying their structure to adhere to domain-specific constraints (Rudin, 2019). However, recent methods such as concept whitening (Chen et al., 2020) and interpretable CNNs (Zhang et al., 2018) show that performance need not be sacrificed for interpretability.

SCENE-Net V2 advances this field by embedding intrinsic geometric interpretability directly into its architecture, mitigating the need for human interpretation biases. By leveraging geometric priors and encoding them through functional observers whose parameters are learned during training, SCENE-Net V2 offers mechanistic insights into its decision-making process.

2.3. GENEO-based Models

The theoretical framework for Group Equivariant Non-Expansive Operators (Bergomi et al., 2019; Cascarano et al., 2021; Conti et al., 2022), although recently developed, has already led to models that demonstrate exceptional performance and efficiency. These GENEO-based models leverage task-specific prior knowledge, which can encompass geometric properties (Lavado et al., 2023; Bergomi et al., 2019), physico-chemical characteristics (Bocchi et al., 2022), and other relevant data attributes. In their seminal work, Bergomi et al. (2019) applied GENEOs to the MNIST dataset, utilizing Gaussian mixture kernels to ensure equivariance under isometric transformations, such as reflections, translations, and rotations. Following this, Bocchi et al. (2022) harnessed GENEOs to detect druggable protein pockets, creating eight GENEOs that encode geometric, physical, and chemical properties of proteins. More recently, Lavado et al. (2023) used GENEOs to characterise the geometric features of pole-shaped structures, enabling the detection of power line supporting towers, an essential task for inspection of power grids. SCENE-Net V2 advances the study of GENEO-based models by addressing the challenge of multiclass 3D semantic segmentation with a novel strategy. Unlike previous approaches that used GENEOs tailored for specific tasks, which significantly restricted their applicability, our model employs operators with general geometric priors. These operators are not designed to detect specific objects in 3D scenes; instead, they represent simple shapes that can be combined to form more complex ones, enabling the detection of various 3D elements. This flexibility allows SCENE-Net V2 to effectively handle a broader range of tasks in 3D semantic segmentation. Our predecessor, SCENE-Net (Lavado et al., 2023), served as a proof-of-concept for GENEO-based models applied to 3D data. It is a fully white-box model with one layer and 11 trainable parameters, specifically designed to detect pole-like structures. SCENE-Net V2 is a deep neural network with more than 500 interpretable parameters and boasts of more geometric priors with several degrees of freedom to adjust to different types of 3D elements. After the feature extraction step, we employ a simple black-box classifier to segment the 3D elements in the input. Thus, SCENE-Net V2 can be seen as a gray-box model that is a sweet-spot between general object identification and transparency.

2.4. Group Equivariant Convolutional Methods

Equivariant architectures have been explored using group convolutions (G-convolutions) (Cohen & Welling, 2016) and formalized in (Cohen et al., 2019). This work focuses on generalizing the conventional translation-equivariant convolution to arbitrary groups via parameterized kernel functions. For instance, Cohen & Welling (2016) show two particular cases of G-convolutions by instantiating G as the *p4* group consisting of all compositions of translations and rotations by 90 degrees about any center of rotation and the *p4m* group, which generalizes the previous group by also considering reflections. The GENEO theoretical framework (Bergomi et al., 2019; Cascarano et al., 2021) can be regarded as an extension of the work of Cohen et al. (2019) as it lays the foundation for general group equivariant operators applied to topological spaces. These operators degenerate into the convolution operator and the Euclidean space where the G-convolutions are defined. Indeed, the convolution operator itself is a GENEO. Bergomi et al. (2019) emphasize that the restriction to certain operator families and equivariance with respect to interpretable transformations (i.e., translation) are key aspects for the architecture's effectiveness. Unlike Cohen et al. (2019), the GENEO framework requires non-expansiveness in GENEO-kernels to ensure faster convergence and approximability by a finite set of operators within the same space (Bergomi et al., 2019; Cascarano et al., 2021). However, this is not necessary for defining general G-equivariant operators, Group Equivariant Operators (GEOs) are also encompassed in the framework of Bergomi et al. (2019); Cascarano et al. (2021).

3. Group Equivariant Non-Expansive Operators

Group Equivariant Non-Expansive Operators (GENEOs) form the core of a mathematical framework (Bergomi et al., 2019; Cascarano et al., 2021; Conti et al., 2022) that characterizes machine learning agents as a set of operators acting on input data. These operators extract essential features from the data, similar to how CNN kernels identify important patterns to recognize objects. These agents, or observers, transform data into higher-level representations while adhering to a set of properties defined by a group of transformations. An effective observer transforms data in a way that commutes with these transformations, making it *equivariant* with respect to the transformation group. The framework is grounded on topological data analysis (TDA) to represent data as sets of functions, which, endowed with the topology induced by the L_∞ norm, become topological spaces. Specifically, a dataset is represented by a set Φ of real valued functions defined on a space X, $\varphi\colon X \to \mathbb{R}$. Φ encompasses all admissible measurements on X. For instance, images can be seen as functions assigning RGB values to pixels. This abstraction allows the framework to focus on the measurement space instead of the raw data. To define prior knowledge, we introduce a group G of transformations on Φ, and we assume that if $\varphi \in \Phi$ and $g \in G$ then $\varphi \circ g \in \Phi$. We call the couple (Φ, G) *perception pair*. The group G represents transformations on the input data for which equivariance is enforced. This group G embeds prior knowledge into the GENEO model.

Definition 3.1 (Group Equivariant Non-Expansive Operator (GENEO))**.** Consider two perception pairs (Φ, G) and (Ψ, H) and a homomorphism $T\colon G \to H$. A map $F\colon \Phi \to \Psi$ is a group equivariant non-expansive operator if it is equivariant,i.e.,

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

and non-expansive, therefore,

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

Non-expansivity and convexity are crucial for the practical application of GENEOs in machine learning. When Φ and Ψ are compact, non-expansivity ensures that the space of all GENEOs $\mathcal{F}$ is also compact (Bergomi et al., 2019; Cascarano et al., 2021), enabling any operator to be approximated by a finite set of operators within the same space. Additionally, if Ψ is convex, Bergomi et al. (2019) prove that $\mathcal{F}$ is also convex. This property allows the convex combination of GENEOs into other GENEOs, ensuring an efficient approximation of any operator by a finite set of GENEOs within the same space.

4. SCENE-Net V2 Architecture

In this section, we present the architecture of SCENE-Net V2, highlighting the innovative aspects that distinguish it from its predecessors. We delve into the geometrical shape priors that form the foundation of our model, describe the method for constructing complex observers using these priors, and outline the end-to-end training process that enables SCENE-Net V2 to achieve good performance and intrinsic geometric interpretability in 3D semantic segmentation.

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 $\varphi\colon \mathbb{R}^3 \to \{0, 1\}$, which signals the presence of 3D points in a voxel discretization. Next, the transformed input is fed into a GENEO layer, which comprises multiple GENEOs selected from a parametric family of operators, each defined by a set of trainable shape parameters ϑ (see Figure 2). These GENEOs operate as convolutional operators with kernels designed to capture essential geometric features. Convolution inherently offers equivariance with respect to translations, making it a suitable operation for this task. During training, the shape parameters ϑ of each GENEO are optimized, rather than the kernels themselves, to preserve equivariance throughout the optimization process.

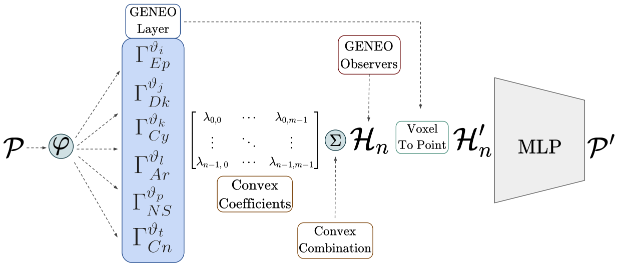
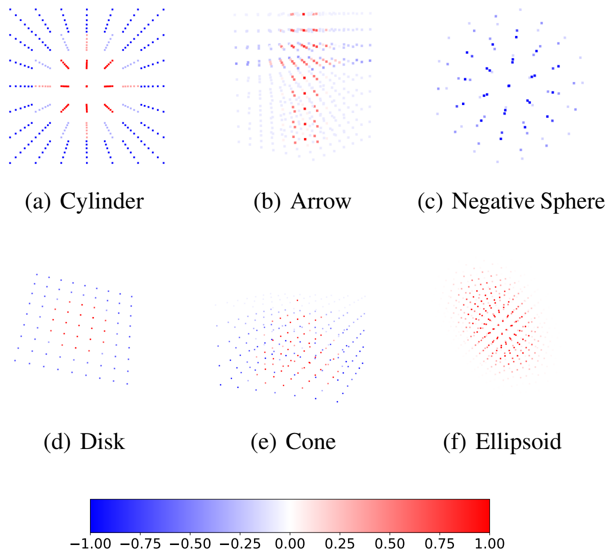

GENEO Layer
$\Gamma^{\vartheta_i}_{Ep}$
$\Gamma^{\vartheta_j}_{Dk}$
$\Gamma^{\vartheta_k}_{Cy}$
$\Gamma^{\vartheta_l}_{Ar}$
$\Gamma^{\vartheta_p}_{NS}$
$\Gamma^{\vartheta_t}_{Cn}$
$\mathcal{P}$
φ
$\begin{bmatrix} \lambda_{0,0} & \cdots & \lambda_{0,m-1} \\ \vdots & \ddots & \vdots \\ \lambda_{n-1,0} & \cdots & \lambda_{n-1,m-1} \end{bmatrix}$
Convex Coefficients
Σ
Convex Combination
GENEO Observers
$\mathcal{H}_n$
Voxel To Point
$\mathcal{H}'_n$
MLP
$\mathcal{P}'$

Figure 2. SCENE-Net V2 architecture: An input point cloud $\mathcal{P}$ is initially measured according to function φ and discretized into a 3D voxel grid. A GENEO Layer with m GENEO kernels then extracts geometric information from the voxel grid. Each operator Γ is based on the convolution operation and is derived from six parametric families of geometric shape priors. Following this, n GENEO observers $\mathcal{H}$ are obtained through a convex combination of the GENEO Layer outputs, with the convex coefficients illustrated by λ. These observers learn to combine features extracted in the GENEO Layer, recognizing complex geometric patterns in the data. Finally, the shape prior features from the GENEO Layer and the GENEO observers are merged via a voxel-to-point transformation, resulting in $\mathcal{H}'$, which is then classified using a Multi-Layer Perceptron (MLP).

The GENEO layer produces a set of operators $\Gamma = \{\Gamma^{\vartheta_j}_j\}^m_{j=1}$ with shape parameters $\vartheta = \vartheta_1, \ldots, \vartheta_m$. These operators are combined through a convex combination with weights $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{n\times m}$, where n is the number of observers and m is the number of kernels, resulting in GENEO observers $\mathcal{H}$. Formally, the observers are defined as:

$$\mathcal{H}_i(x) = \sum_{j=1}^{m} \Lambda_{ij} \Gamma^{\vartheta_j}_j(\varphi)(x) \tag{3}$$

Since the convex combination of GENEOs is also a GENEO (Bergomi et al., 2019), each $\mathcal{H}_i$ maintains the equivariance properties of the individual operators $\Gamma^{\vartheta_j}_j$. These observers analyze the 3D input scenes, detecting the various geometric properties encoded in Γ and learn how to combine them through Λ_i. The convex coefficients Λ, illustrated in the architecture, represent the contribution of each operator $\Gamma^{\vartheta_j}_j$ to each observer $\mathcal{H}_i(x)$, providing intrinsic interpretability to the model. Following the GENEO layer and observer combination, the shape prior features and GENEO observers' outputs are merged through a voxel-to-point transformation. This combined output $\mathcal{H}'$ is then classified using a Multi-Layer Perceptron (MLP), which assigns semantic labels to each point in the 3D point cloud.


(a) Cylinder (b) Arrow (c) Negative Sphere

(d) Disk (e) Cone (f) Ellipsoid

Figure 3. GENEO kernels discretized in a voxel grid.

4.2. Building GENEOs from Geometric Priors

GENEOs act on functions, transforming them to remain equivariant to a specific group of transformations. Our GENEOs act upon Φ, the topological space of admissible functions $\varphi\colon \mathbb{R}^3 \to \{0,1\}$ representing the measurements done on the voxel discretization of the space where $\mathcal{P}$ is located. 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, a GENEO Γ^ϑ transforms φ into a new function that detects sections in the input point cloud demonstrating the properties of a geometric prior g (where g is a kernel function) while preserving the geometry of the 3D scene:

$$\Gamma^\vartheta\colon \Phi \to \Psi, \qquad \psi = \Gamma^\vartheta(\varphi)$$
$$\psi(x) = \int_{\mathbb{R}^3} \tilde{g}(y)\varphi(x-y)dy. \tag{4}$$

Here Ψ is a new functional space that represents $\mathcal{P}$ with functions $\psi\colon \mathbb{R}^3 \to [0,1]$, and $\tilde{g}$ defines a normalized version of the geometric prior g. The kernel g 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 geometrical behaviour and punish those which do not. Thus, $\psi(x)$ assumes positive values for 3D points that exhibit the desired geometric properties, whereas negative values discourage shapes that do not fall under the definition of g. This leads to the detection of a set of geometrical properties or structures emulating the result of the processing of the point cloud by visual inspection by a human observer.

4.2.1. Cylinder, Arrow and Negative Sphere

The Cylinder, Arrow and Negative Sphere geometric priors were introduced in (Lavado et al., 2023) to fully describe power line supporting towers (i.e., pole-like structures). Still, they define signature shapes found regularly in 3D point clouds:

The **Cylinder GENEO** is rotationally equivariant around the *z-axis* and translationally equivariant within the *xy* plane, forming a cylinder shape. Mathematically, it is defined as:

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

Here, $z(\cdot)$ represents the projection of a 3D point onto the *xy* plane, and c denotes the center of the cylinder. The shape parameters $\vartheta_{Cy} = [r, \sigma]$ determine the radius of the cylinder and the spread of the smoothing Gaussian. Figure 3(a) illustrates the kernel g_{Cy} discretized in a voxel grid.

The **Arrow GENEO** is used to identify structures resembling an arrow, specifically a cone atop a cylinder, which can represent certain architectural features. This operator is defined as:

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

where $\pi_i(x)$ is a projection function of ith elements of the input vector. This operator's shape parameters $\vartheta_{Ar} = [r, \sigma, h, r_c, \beta]$ define the dimensions and inclination of the cone and cylinder, allowing it to capture complex vertical structures. Figure 3(b) illustrates the kernel g_{Ar} discretized in a voxel grid.

The **Negative Sphere GENEO** was originally employed in (Lavado et al., 2023) to suppress spherical shapes commonly found in vegetation, such as trees, thus reducing false positives in the detection of non-target structures. It is defined as:

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

With the shape parameters $\vartheta_{NS} = [r, \sigma, \omega]$, this operator penalizes spherical geometries, effectively distinguishing target structures from vegetation. In our work, we utilize this GENEO but no longer enforce the weight ω to be positive, thus, it can be used to detect or diminish spherical structures in the data. Figure 3(c) illustrates the kernel g_{NS} discretized in a voxel grid.

4.2.2. Disk GENEO

The Disk GENEO encodes a plane surface disk as a geometric prior. It initially provides rotational equivariance around the z-axis, similar to the Cylinder GENEO. However, the disk definition includes two additional learnable angles, ζ

and β, which correspond to rotations around the *x* and *y* axes, respectively. This flexibility allows the Disk GENEO to provide rotational equivariance to the normal vector of the learned angles during training. Such adaptability is useful for detecting both ground surfaces and building walls. The disk prior is defined as follows:

$$g_{Dk}(x) = \begin{cases} e^{\frac{-1}{2\sigma^2}(\|z(R_{\zeta,\beta}(x))-z(c)\|^2-r^2)^2} & \pi_3(R_{\zeta,\beta}(x)) = h \\ 0 & \text{otherwise} \end{cases}$$

Here, g_{Dk} represents a smoothed disk pattern. $R_{\zeta,\beta}(a)$ denotes the rotation of a by angles ζ and β around the *x* and *y* axes, respectively. The parameter *c* denotes the center of the disk, r is the radius of the disk, and σ controls the spread of the smoothing Gaussian. The shape parameters $\vartheta_{Dk} = [\zeta, \beta, \sigma]$ allow the disk to adapt to different orientations and to regulate its shape. Finally, h defines the height at which the disk lies in the *z* dimension. Figure 3(d) illustrates the kernel g_{Dk} discretized in a voxel grid.

4.2.3. CONE GENEO

The cone prior is designed to capture small scene elements that may vary in shape, such as small vegetation and tree tops that do not align with a spherical crown. Its learnable base radius allows it to adapt to various objects in 3D scenes. The Cone GENEO provides equivariance w.r.t. rotations on the *z*-axis. Mathematically, it is defined as:

$$g_{Cn}(x) = e^{\frac{-1}{2\sigma^2}(\|z(x)-z(c)\|^2-(r\tan(\beta\pi))^2)^2}$$

With shape parameters $\vartheta = [r, \sigma, \beta]$, where r is the cone radius, σ is the spread of the Gaussian and β is inclination angle of the cone. These parameters enable the Cone GENEO to effectively model a variety of small, conical structures within 3D scenes. Figure 3(e) illustrates the kernel g_{Cn} discretized in a voxel grid.

4.2.4. ELLIPSOID GENEO

This GENEO is a generalization of the negative sphere. By designing a geometric prior with more degrees of freedom w.r.t. its shape, we strive for a better adaptability to different 3D elements. The Ellipsoid GENEO provides rotational equivariance along the axes of its focal points. These are learned during training as the geometric prior is adapted to the input data. The definition of the prior goes as follows:

$$g_{El}(x) = \omega e^{-\frac{1}{2}\left((x-c)^T\Sigma^{-1}(x-c)\right)},$$

where Σ is the covariance matrix defined by the radii along the *x*, *y*, and *z* axes, given by $\Sigma = \text{diag}(a^2, b^2, c^2)$. Here, ω is the scaling factor, c is the center of the ellipsoid, and a, b, and c are the radii along the principal axes. The shape parameter vector is $\vartheta = [a, b, c, \sigma, \omega]$. Figure 3(f) illustrates the kernel g_{El} discretized in a voxel grid.

4.3. Optimizing SCENE-Net V2

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

$$\begin{aligned} \underset{\Lambda,\vartheta}{\text{minimize}} \quad & \mathop{\mathbb{E}}_{(X,y)\sim\mathcal{D}} [\mathcal{L}_{seg}(\Lambda, \vartheta; X, y)] \\ \text{s.t.} \quad & \Lambda \in \Delta^{(m-1)\times n}, \quad \vartheta \in \mathbb{R}_+^T, \end{aligned} \tag{5}$$

where $\Delta^{(m-1)\times n}$ denotes the $(m-1)$-dimensional simplex for each of the n observers, ensuring that the weights Λ form a convex combination. $\mathbb{R}_+^T$ represents the non-negative orthant for all shape parameters ϑ. These constraints are crucial, as parameters such as the radius of a cylinder must remain non-negative to be meaningful.

The segmentation loss $\mathcal{L}_{seg}$ is defined as a weighted cross-entropy term:

$$\mathcal{L}_{seg}(\Lambda, \vartheta; X, y) = f_w(\alpha, \epsilon; y)\text{CE}\Big(\underset{\Lambda,\vartheta}{\mathcal{M}}(X), y\Big) \tag{6}$$

where $f_w(\alpha, \epsilon; y)$ denotes 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 $\mathcal{D}$. The hyperparameter α emphasizes the weighting scheme, while ϵ ensures positive weights by acting as a small positive number. CE calculates the cross entropy loss between our model's prediction $\underset{\Lambda,\vartheta}{\mathcal{M}}(X)$ and the class labels y. To facilitate optimization, we reparametrize Λ to satisfy the simplex constraint by setting $\Lambda_{m,j} = 1 - \sum_{i=1}^{m-1} \Lambda_{i,j}$ for each observer j. This reduces the problem to:

$$\begin{aligned} \underset{\Lambda,\vartheta}{\text{minimize}} \quad & \mathop{\mathbb{E}}_{(X,y)\sim\mathcal{D}} [\mathcal{L}_{seg}(\Lambda, \vartheta; X, y)] \\ \text{s.t.} \quad & \vartheta \in \mathbb{R}_+^T, \quad \Lambda \in \mathbb{R}_+^{m\times n}, \end{aligned} \tag{7}$$

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

$$\underset{\Lambda,\vartheta}{\text{minimize}} \quad \mathop{\mathbb{E}}_{(X,y)\sim\mathcal{D}} [\mathcal{L}_{seg}(\Lambda, \vartheta; X, y)] + \Omega(\Lambda, \vartheta), \tag{8}$$

where the regularization term $\Omega(\Lambda, \vartheta)$ is the combination of a negativity penalty and the Elastic Net (Zou & Hastie, 2005) penalty and is defined as follows:

$$\begin{aligned} \Omega(\Lambda, \vartheta) = \rho_l & \left(\sum_{j=1}^{n} \sum_{i=1}^{m} h(\Lambda_{i,j}) + \sum_{i=1}^{T} h(\vartheta_i) \right) + \\ \rho_t & \left(\eta \sum_{j=1}^{n} \sum_{i=1}^{m} \|\Lambda_{i,j}\|_1 + (1-\eta) \sum_{j=1}^{n} \sum_{i=1}^{m} \|\Lambda_{i,j}\|_2^2 \right) \end{aligned} \tag{9}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2^2$ represent the L1 and L2 norms, respectively. The hyperparameter $\eta \in [0, 1]$ controls the trade-off between L1 and L2 regularization, while ρ_l and ρ_t are the regularization coefficients of the two penalties. Lastly, $h(x) = \max(0, -x)$ penalizes negative values. The use of the Elastic Net penalty is crucial for building complex observers. It enables the pruning of unwanted behaviors, such as some observers focusing solely on a single GENEO kernel or having equally distributed convex weights. By promoting sparsity and small weights in the convex coefficient matrix Λ, SCENE-Net V2 is able to focus on useful geometric priors and optimally adapt them to the input data.

Overall, this formulation effectively balances data fidelity, as measured by the segmentation loss, with the theoretical guarantees of convexity and non-negativity required by the GENEO framework, ensuring robust and meaningful shape parameter learning throughout the training process.

5. Experiments

TS40K Dataset. We evaluate our models on the TS40K dataset (Lavado et al., 2024), a unique outdoor 3D point cloud dataset covering over 40,000 kilometers of the European electrical transmission system in rural areas. This dataset enhances inspection processes, crucial for preventing power outages and forest fires. With drone-based LiDAR scans replacing traditional inspections, TS40K offers unique 3D data properties: high point-density, absence of occlusion, and homogeneous point-density, unlike self-driving benchmarks. Additionally, inspection-based annotations introduce noise and mislabeled points, reflecting real-world conditions, and extreme class imbalance poses a significant challenge, with transmission system-related classes being underrepresented. The data properties of TS40K result in extremely detailed 3D elements, making this dataset especially suited to assess the use of geometric priors.

5.1. Results and Analysis

Baselines. We conduct a comprehensive evaluation of SCENE-Net V2 by exploring several variants of the model. Specifically, we examine the impact of different numbers of GENEO kernels, observers, GENEO layers, and kernel sizes. Additionally, we perform an ablation study to assess the influence of the geometric priors used. To provide a robust comparison, we benchmark SCENE-Net V2 against a Convolutional Neural Network (CNN) with a similar architecture and the same base operator (i.e., convolution) for feature extraction. The primary distinction lies in the kernel initialization: while CNN kernels are initialized randomly, our model's kernels are also initialized randomly but are derived from a specific family of operators parameterized by shape parameters. This unique characteristic makes SCENE-Net V2 independent of kernel size, meaning the number of parameters remains constant regardless of the kernel size used to discretize the GENEO kernels. This is particularly advantageous in 3D applications, where traditional CNNs face exponentially increasing memory and computational costs with larger kernel sizes. Furthermore, the classifier architecture is consistent across all experiments. For a fair comparison with state-of-the-art models, we evaluate SCENE-Net V2 against leading benchmarks in 3D semantic segmentation. In this evaluation, we also consider model sizes to account for the parameter efficiency.

Table 1. 3D Semantic segmentation results of TS40K test set. Only methods that report the parameter count were included. Here we report mean Intersection over Union (mIoU %), number of parameters, parameter efficiency $\frac{\text{mean IoU}}{\text{log \#Parameters}}$ and whether the model is interpretable.

Method	mIoU	#Parameters (M)	Parameter Efficiency	Is Interpretable?
PointNet (Qi et al., 2017a)	44.58	0.40	7.96	No
PointNet++ (Qi et al., 2017b)	46.90	1.48	7.60	No
KPConv (Thomas et al., 2019)	57.58	14.9	8.03	No
RandLA-Net (Hu et al., 2020)	16.76	1.24	2.75	No
Point Transformer V1 (Wu et al., 2022)	62.67	12.8	8.81	No
Point Transformer V2 (Wu et al., 2023a)	**65.58**	46.2	8.56	No
CNN Baseline	41.69	0.26	7.69	No
SCENE-Net V2 (Ours)	45.54	**0.24**	8.46	**Yes**
SCENE-Net V2 + CNN (Ours)	**50.21**	0.26	**9.27**	**Yes**

Performance of SCENE-Net V2. Table 1 compares various 3D semantic segmentation benchmarks on the TS40K test set. While Point Transformer V2 (Wu et al., 2023a) achieves the highest mean IoU (65.58%), it requires a substantial 46 million parameters, more than three times the parameter count of its predecessor, Point Transformer V1 (Wu et al., 2022), to achieve less than a 3% increase in mIoU. The parameter efficiency metric highlights the diminishing returns of enlarging model sizes to enhance performance. Our proposed model, SCENE-Net V2, achieves a competitive mIoU of 45.54% and a leading parameter efficiency of 8.46, outperforming models such as PointNet (Qi et al., 2017a), RandLA-Net (Hu et al., 2020), and the CNN baselines with a similar architecture. SCENE-Net V2 stands out for being the only model to offer intrinsic interpertability through geometric priors while achieving competitive performance. SCENE-Net V2 has a total of 240K parameters, with just 540 dedicated to the GENEO feature extraction step. The remaining parameters are part of the MLP classifier. In comparison, the CNN baseline's feature extraction comprises 21.4K parameters and its mIoU is 41.69%, 4% less than our model. We also explore the use of SCENE-Net V2 as a feature extraction tool from 3D point clouds for black-box models. As a proof of concept, we add a GENEO feature extraction step, composed of just 540 meaningful parameters, to the CNN baseline. This results in an improved mean IoU of 50.21%, representing a 8.52% performance boost for the CNN baseline and a parameter efficiency increase of 1.58, recording the highest value among all benchmarks.

Kernel Size. Table 2 provides a comprehensive overview of the performance of SCENE-Net V2 on the TS40K dataset

Table 2. Performance of SCENE-Net V2 on the TS40K test set with different kernel sizes. We report mean IoU (mIoU %) and per-class IoU (%) scores.

Kernel Size (z, x, y)	mIoU	Ground	Low Vegetation	Medium Vegetation	Power Line Supporting Tower	Power Line
(3, 3, 3)	28.09	52.22	9.74	25.13	19.89	33.45
(5, 5, 5)	33.15	52.78	14.13	26.15	20.76	51.94
(7, 7, 7)	36.08	61.44	13.78	28.41	**34.57**	42.19
(9, 9, 9)	37.08	59.53	11.08	28.24	20.26	66.31
(9, 5, 5)	37.71	58.17	12.34	27.09	19.75	71.22
(9, 7, 7)	32.68	57.77	12.29	27.53	22.85	42.97
(12, 12, 12)	35.29	57.30	9.21	29.34	22.98	57.62
(12, 5, 5)	**45.54**	**64.49**	**17.84**	**34.79**	21.92	**88.66**
(12, 7, 7)	32.62	53.66	12.94	28.92	17.24	50.36
(5, 9, 9)	32.04	60.36	11.45	29.24	16.28	42.86
(5, 12, 12)	33.47	55.49	14.53	28.23	16.98	52.12

Table 3. Performance of SCENE-Net V2 on the TS40K test set with different numbers of observers and a kernel size of (7, 7, 7). We report mean IoU (mIoU %) and per-class IoU (%) scores.

Number of Observers	mIoU	Ground	Low Vegetation	Medium Vegetation	Power Line Supporting Tower	Power Line
8	32.87	59.30	**18.57**	26.74	15.93	**43.82**
16	**35.79**	**61.44**	12.89	**28.65**	**34.52**	41.47
32	32.17	60.53	14.12	26.95	16.74	42.51
64	28.64	57.08	10.57	26.32	11.94	37.29
128	29.39	55.76	10.92	27.16	13.58	39.53

with various kernel sizes. The analysis reveals that the kernel size (12, 5, 5) achieves the highest mean IoU of 45.54%, indicating the best overall segmentation performance. Additionally, it shows the best per-class IoU in all classes except for power line supporting towers. Kernel sizes (9, 5, 5) and (9, 9, 9) also perform notably well, with mean IoUs of 37.71% and 37.08%, respectively. In contrast, the smallest kernel size (3, 3, 3) records the lowest mean IoU of 28.09%. This suggests that smaller kernels may not capture sufficient information for accurate segmentation in this dataset. The better performance of larger, narrower kernels (e.g., (12, 5, 5)) could be attributed to the effective discretization of the geometric priors in 3D kernels. For example, a rotating disk prior might be poorly discretized in a (3, 3, 3) kernel, with adjustments to its rotating angles and Gaussian spread having little to no impact on the final kernel.

Number of Observers. In Table 3 we study the performance of SCENE-Net V2 with varying numbers of observers. The model with 16 observers achieves the highest mean IoU at 35.79%, demonstrating superior segmentation performance. However, increasing the number of observers beyond 16 leads to a decline in performance. This trend suggests that additional observers might introduce redundancy and contribute to overfitting.

Table 4. Performance of SCENE-Net V2 on the TS40K test set with different GENEO kernel counts. We report mean IoU (mIoU %) and per-class IoU (%) scores. The number of reported GENEO kernels is for each type of geometric prior.

GENEO Kernel Count Per Geometric Prior	mIoU	Ground	Low Vegetation	Medium Vegetation	Power Line Supporting Tower	Power Line
4	30.73	55.45	10.73	27.86	25.39	34.21
8	**37.57**	58.87	12.76	28.01	17.98	**70.21**
16	35.79	**61.44**	**12.89**	28.65	34.52	41.47
32	31.89	59.90	11.14	28.27	20.59	39.58
64	33.37	60.95	13.75	**29.06**	21.13	41.98
128	29.05	49.24	11.28	25.57	**37.01**	22.13

Table 5. Performance of SCENE-Net V2 on the TS40K test set with different numbers of GENEO layers. We report mean IoU (mIoU %) and per-class IoU (%) scores.

GENEO Layer Count	mIoU	Ground	Low Vegetation	Medium Vegetation	Power Line Supporting Tower	Power Line
(16)	**35.79**	**61.44**	**12.89**	**28.65**	**34.52**	**41.47**
(8, 16)	28.10	48.79	11.26	25.24	21.73	33.49
(8, 16, 32)	28.67	58.43	10.97	26.53	11.84	35.59
(8, 16, 32, 64)	27.95	51.83	11.87	25.79	20.36	29.91

Number of GENEO kernels. In Table 4, we analyze the performance of SCENE-Net V2 on the TS40K dataset with different GENEO kernel counts. The model with 8 GENEO kernels achieves the highest mean IoU at 37.57%, indicating the best overall segmentation performance. On the other hand, the model with 128 GENEO kernels records the lowest mean IoU of 29.05%, suggesting that an excessively large number of kernels may lead to reduced performance.

Going Deeper? Contrary to traditional deep learning methodologies, GENEO-based models consist of a singular feature extraction layer. While conventional architectures often leverage multiple layers to progressively extract hierarchical features, GENEO models perform feature extraction with a single layer. Moreover, the convex combination of such features does not introduce new feature extraction. In Table 5, we assess how varying numbers of GENEO layers impact the performance of SCENE-Net V2. Unlike traditional models, increasing the number of layers doesn't enhance performance in GENEO-based models.

Ablation Study. In Table 6, we study the impact of ablating different geometric priors on the performance of SCENE-Net V2. The results show that the absence of the cylinder prior results in the most significant drop in mean IoU. Notably, the IoU for power line supporting towers is at its lowest without the cylinder prior, highlighting the critical role of the Cylinder GENEO in extracting relevant geometric information from raw point clouds. Among the ablated models, the removal of the Arrow GENEO results in the smallest performance decrease. This suggests that the features captured by the Arrow GENEO can be effectively approximated by the cone and cylinder priors. Finally, the performances without the negative sphere and ellipsoid priors are very similar. This similarity is expected, as the ellipsoid is a generalization of the sphere, indicating that both priors capture comparable geometric properties.

5.2. Interpretability Analysis of GENEO Observers

SCENE-Net V2 consists of two phases: transparent feature extraction and classification. Firstly, the input point

Table 6. Performance of SCENE-Net V2 on the TS40K test set with different geometric priors ablated and 8 observers. We report mean IoU (mIoU %) and per-class IoU (%) scores.

Geometric Prior Ablation	mIoU	Ground	Low Vegetation	Medium Vegetation	Power Line Supporting Tower	Power Line
No Cylinder	26.96	50.34	7.50	25.86	11.19	39.91
No Negative Sphere	33.03	55.63	6.17	29.36	**43.23**	30.75
No Arrow	36.83	**63.01**	**18.41**	**31.32**	20.01	51.38
No Disk	27.93	40.31	7.83	25.32	36.44	29.75
No Cone	34.59	62.04	6.42	29.17	18.04	57.26
No Ellipsoid	33.72	51.18	9.42	27.35	40.71	39.92
All GENEOs	**37.57**	58.87	12.76	28.01	17.98	**70.21**

cloud is processed using geometric priors to identify key 3D shapes. These priors, defined by meaningful shape parameters are inherently interpretable. They are then combined into **observers** through convex combinations, creating more complex feature extraction outputs. The convex coefficients Λ indicate the contribution of each prior, ensuring the transparency of observers. The second phase involves classifying the features extracted by the GENEO observers. Compared to the simple white-box architecture of SCENE-Net (Lavado et al., 2023), our model is a gray-box model due to its larger size and use of a black-box classifier. We trade-off full interpretability for general application. Even though not all parameter hold meaning, we can link the classification of 3D points to specific priors and also adjust the network accordingly. For example, in Figure 4(a) an observer processes a TS40K sample, and by examining its convex coefficients we identify the most significant prior for this observer: a disk with minimal rotation and a small radius. Further analysis into the activations of the MLP classifier indicate that the observer is crucial in the segmentation of power lines (seen in Figure 4(b)). Thus, the transparency of GENEO-based models allows for a meaningful analysis and evidence-based adjustments to a model's architecture. For instance, identifying the disk prior's importance in detecting power line towers explains why noise above the grid is misclassified as a power line, as seen in the top row of Figure 4.

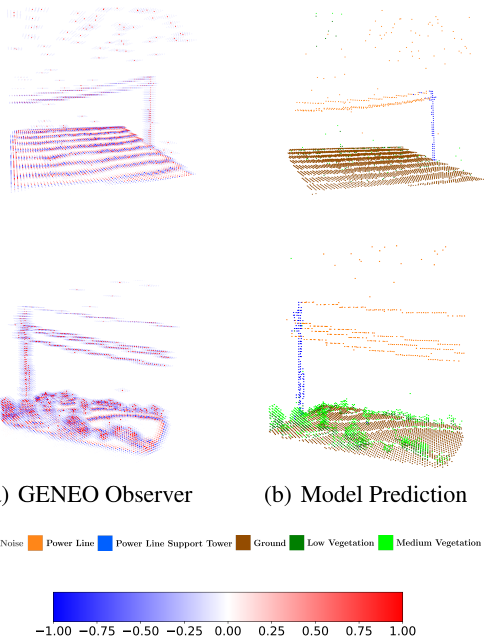


Figure 4. Visualizing the inner workings of SCENE-Net V2: A GENEO observer (a) is constructed through the convex combination of various geometric priors. By examining the convex coefficients, we identify that the most influential prior is a disk with minimal rotation and a radius of 0.02. In SCENE-Net V2, these observers are then processed by a standard MLP classifier, which generates the model predictions (b).

6. Conclusions

In this paper, we introduced SCENE-Net V2, the first gray-box model for multiclass 3D scene understanding. By leveraging Group Equivariant Non-Expansive Operators (GENEOs), SCENE-Net V2 incorporates fundamental geometric priors in its feature extraction step. We addressed the limitations of the predecessor model, SCENE-Net (Lavado et al., 2023), by significantly expanding its scope of application from pole-like structures to potentially any dataset with detailed 3D elements. Our results demonstrate that SCENE-Net V2 achieves a competitive performance with the lowest parameter count and that using a GENEO-based feature extraction step in black-box models leads to a significant increase in performance with just 540 meaningful parameters. For future work, we will further explore the use of GENEO-based architectures as geometric feature extraction tools in 3D scene understanding.

Acknowledgements

We thank F. Metelo, P. Valdeira, R. Belo and S. Kumar for the interesting discussions. 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" Nº C644943391-00000051/40 co-funded by Recovery and Resilience Plan and NextGeneration EU, and TaRDIS Horizon Europe (101093006). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia
NextGenerationEU

References

Barbiero, P., Ciravegna, G., Giannini, F., Lió, P., Gori, M., and Melacci, S. Entropy-based logic explanations of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6046–6054, 2022.

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

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

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

Chen, Z., Bei, Y., and Rudin, C. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Cheng, R., Razani, R., Taghavi, E., Li, E., and Liu, B. 2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 12547–12556, 2021.

Choy, C., Gwak, J., and Savarese, S. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3075–3084, 2019.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Cohen, T. S., Geiger, M., and Weiler, M. A general theory of equivariant cnns on homogeneous spaces. *Advances in neural information processing systems*, 32, 2019.

Conti, F., Frosini, P., and Quercioli, N. On the construction of group equivariant non-expansive operators via permutants and symmetric functions. *Frontiers in artificial intelligence*, 5:786091, 2022.

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

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., and Bennamoun, M. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., and Markham, A. 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, 2020.

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

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

Lavado, D., Soares, C., Micheletti, A., Bocchi, G., Coronati, A., Silva, M., and Frosini, P. Low-resource white-box semantic segmentation of supporting towers on 3d point clouds via signature shape identification. *arXiv preprint arXiv:2306.07809*, 2023.

Lavado, D., Soares, C., Micheletti, A., Santos, R., Coelho, A., and Santos, J. Ts40k: a 3d point cloud dataset of rural terrain and electrical transmission system. *arXiv preprint arXiv:2405.13989*, 2024.

Lawin, F. J., Danelljan, M., Tosteberg, P., Bhat, G., Khan, F. S., and Felsberg, M. Deep projective 3d semantic segmentation. In *Computer Analysis of Images and Patterns: 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part I 17*, pp. 95–107. Springer, 2017.

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

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., and Chen, B. PointCNN: Convolution on x-transformed points. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Lyu, Y., Huang, X., and Zhang, Z. Learning to segment 3d point clouds in 2d image space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12255–12264, 2020.

Meng, H.-Y., Gao, L., Lai, Y.-K., and Manocha, D. Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8500–8508, 2019.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. 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, 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 information processing systems*, 30, 2017b.

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

Ribeiro, M. T., Singh, S., and Guestrin, C. " 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, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Steininger, M., Kobs, K., Davidson, P., Krause, A., and Hotho, A. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 945–953, 2015.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., and Guibas, L. J. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 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 Information Processing Systems*, 35:33330–33342, 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. *arXiv preprint arXiv:2312.10035*, 2023a.

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, 2023b.

Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., and Li, Z. 2DPASS: 2D Priors Assisted Semantic Segmentation on LiDAR Point Clouds. *arXiv e-prints*, art. arXiv:2207.04397, July 2022.

Yang, Z. and Wang, L. Learning relationships for multi-view 3d object recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7505–7514, 2019.

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

Zhang, Y., Zhou, Z., David, P., Yue, X., Xi, Z., Gong, B., and Foroosh, H. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9601–9610, 2020.

Zhou, Y. and Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4490–4499, 2018.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.