

R-SHAP: Region Based SHAP for 3D Point Cloud Explainability

Muhammad Shoaib Sarwar^[0000–0001–8951–4320], Marc F. Harinck^[0009–0000–8042–8071], and Faizan Ahmed^[0000–0002–2760–6892]

Department of Computer Science, University of Twente, Enschede, Netherlands
shoaib.sarwar@utwente.nl;m.f.harinck@student.utwente.nl;faizan.ahmed@utwente.nl

Abstract. SHapley Additive exPlanations have become a popular tool for feature attribution in machine learning. However, their application to 3D point clouds faces significant challenges because evaluating 2^N point coalitions is computationally expensive for clouds with thousands of points. Existing point-wise approximations use Monte Carlo sampling, often producing explanations that focus on only a few highly sensitive points. These explanations fail to capture the structural information that defines a 3D shape. We introduce R-SHAP, a region-based explanation framework that formulates Shapley attribution as a cooperative game involving M disjoint regions of the point cloud. R-SHAP integrates three ideas. First, a rotation-equivariant Farthest Point Sampling and Voronoi assignment divides the point cloud into balanced regions. Second, a Geometric Interpolation Reference fills absent regions with a Gaussian-weighted average of their nearest present neighbors and maintains perturbations close to the data manifold. Third, a Kernel SHAP weighted regression determines region attributions using a small number of forward passes through the classifier. We implement R-SHAP with the PointNet classifier on the ModelNet10 dataset. The resulting attributions are quantitatively faithful and explainable. The mean faithfulness reaches 0.310 compared to 0.033 for a random ranking. Removing the highest attributed region reduces the predicted logit by 2.45 times as much as removing a random region.

Keywords: Explainable AI · SHAP · Point Cloud Classification · PointNet · Region Attribution.

1 Introduction

Within the field of explainable AI (XAI), SHapley Additive exPlanations (SHAP) has become a widely recognized and mathematically robust method. SHAP provides a rigorous, stable, and equitable allocation of feature importance across model predictions. Originating from cooperative game theory [20,11], SHAP offers both local and global explainability [8]. However, a significant limitation of exact SHAP is its reliance on evaluating all possible feature coalitions, a process that scales exponentially with the number of features as 2^n combinations [8].

To address this immense computational complexity, researchers have developed point-wise approximation methods. Frameworks such as 3D-SVAM [9] and SPAX [6] employ Monte Carlo sampling to estimate Shapley value point attributions. However, point-wise SHAP approaches reveal neural networks’ susceptibility to data corruption [25], indicating that 3D classifiers frequently rely on a limited set of highly sensitive spatial points [25]. Evaluating individual points in isolation does not adequately capture the structural information essential to 3D shape representation [16]. Consequently, recent Shapley based XAI methods have their shifted focus toward broader spatial

segments, partitioning sub regions using farthest point sampling [21,16] and emphasizing regions rather than individual points.

This work introduces R-SHAP, a novel region-based explanation framework for point clouds utilizing the PointNet classifier [3]. R-SHAP employs a rotation-equivariant region segmentation algorithm that uses Farthest Point Sampling (FPS) and Voronoi assignment to group individual points into eight distinct regions. A Geometric Interpolation Reference (GIR) mechanism is implemented to ensure that structural modifications do not artificially influence the classifier’s predictions. Consequently, R-SHAP delivers macro-level feature attributions for regions of points, addressing the computational inefficiency of point-wise Shapley value attribution while preserving spatial complexity [16].

2 Related Work

XAI is increasingly recognized as essential for interpreting deep learning models [8], thereby supporting accountability, trust, and fairness across diverse domains. Within the XAI domain, SHAP has gained prominence as a method that provides mathematically rigorous explanations based on cooperative game theory [20].

Atik et al. [2] applied SHAP to quantitatively explain aerial photogrammetry classifications by extracting 19 engineered geometric and spectral features, such as height, planarity, and RGB values, for individual points. Although this approach is effective for feature analysis in ensemble models like Random Forest and LightGBM, extending these granular, point-wise methods to large, dense point clouds presents significant computational challenges [2]. Tian et al. [25] employed point-wise Shapley values to demonstrate that neural networks often rely disproportionately on a small subset of high-contributing key points. Their findings indicate that pruning these highly sensitive spatial features leads to model misclassification, thereby revealing the pronounced vulnerability of 3D classifiers to data corruption [25].

Xia et al. [27] found that using raw predicted probabilities, instead of binary classification accuracy, captures more nuanced informational entropy. This probabilistic utility function reduces the diminishing returns of marginal contributions and demonstrates improved computational stability [27]. Kuroki and Yamasaki [9] introduced 3D-SVAM, a point-wise Shapley approach that employs a Monte Carlo approximation to efficiently estimate spatial feature attributions for 3D object detection in LiDAR point clouds, resulting in substantial computational savings. Additionally, optimizing the underlying utility function can further enhance the stability of Shapley value computations.

Although SHAP offers both local and global explanations [8], a significant limitation of point-wise SHAP is its dependence on calculating all possible feature coalitions (2^n combinations) or employing approximations such as the Monte Carlo method [9]. In either case, the exponential growth of the feature set renders the computation of exact Shapley values for high-dimensional data, such as 3D point clouds, highly resource-intensive [8]. To address the limitations of point-wise attribution, recent approaches have focused on evaluating broader spatial segments. Günen [4] applied a modified DBSCAN density-based clustering algorithm to segment complex 3D rock geometries into subsets. By implementing TreeSHAP on these subsets, the study effectively assessed the influence of macro-regional geometric features, identifying normal vectors and verticality as the most significant factors for classification [4].

Early regional XAI approaches employed techniques such as FPS to partition 3D space and mitigate the exponential complexity of Shapley value computations [21]. However, Mulawade et

al. [16] showed that naive, purely Euclidean implementations of FPS produce region segments based solely on mathematical criteria, disregarding the physical structure of objects in the ModelNet 10 dataset [26]. Further validation is required to explain these segmented regions meaningfully.

Beyond Shapley attribution, point cloud explainability methods can be grouped by their focus. One group relies on gradients, such as Grad-CAM [19] and integrated gradients, as in Gupta et al. [5] and Levi and Gilboa [10]. Another group uses soft masks to identify important points, like PointMask [23]. Surrogate or proxy models offer another approach, as shown by BubbleX [12] and the work of Tan and Kotthaus [24]. Some methods also draw on concepts or counterfactuals, such as XPCC [1], counterfactuals for 3D data [7], and semantic segmentation [13]. Reviews by Saranti et al. [18] and Mulawade et al. [15] list these groups and highlight the unique value of perturbation based Shapley methods.

R-SHAP differs from earlier region based Shapley pipelines through three main design choices. FPS and Voronoi partition use pairwise distances, so the method is rotation equivariant, unlike Shen et al. [21]. The Geometric Interpolation Reference fills missing regions with manifold-faithful values instead of the origin or random points, addressing concerns from Mulawade et al. [16] and Tian et al. [25]. R-SHAP also reduces the coalition space from 2^N to 2^M with $M \ll N$, following object regions, unlike the point level sampling in Harinck et al. [6]. These choices show R-SHAP as a clear example of the explainable input framework of Lundberg and Lee [11] using rotation equivariant regions, faithful references, and a coalition space that stays small regardless of the point cloud size.

3 Methodology

R-SHAP explains point-wise Shapley attribution as a cooperative game involving a small set of spatial regions. Fig. 1 provides an overview of the four main stages. These stages are region segmentation, coalition sampling, structural reference filling, and weighted regression for recovering Shapley values. Each component is described in the following sections.

3.1 Shapley Values for a Region Game

The Shapley value [20] is the unique fair allocation in a cooperative game (M, v) that satisfies efficiency, symmetry, the null-player property, and linearity. For player j ,

$$\phi_j(v) = \sum_{S \subseteq \mathcal{M} \setminus \{j\}} \frac{|S|! (M - |S| - 1)!}{M!} [v(S \cup \{j\}) - v(S)], \quad (1)$$

where $\mathcal{M} = \{1, \dots, M\}$. SHAP [11,22] adapts these axioms to model attribution by treating features as players. Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathbb{R}^3$ be a point cloud, f_c the logit of the predicted class c , and $\mathcal{R} = \{R_1, \dots, R_M\}$ a partition of the point indices into M disjoint regions covering all N points. Each region acts as a player. For a coalition $S \subseteq \mathcal{M}$, the present-point index set is $I_S = \bigcup_{j \in S} R_j$ and the absent set $I_{\bar{S}}$ is its complement.

Given a reference mechanism ρ that resamples absent points, the *region value function* is

$$v_f(S) = \mathbb{E} \left[f_c(\mathbf{X}^{(S)}) \right], \quad \mathbf{x}_i^{(S)} = \begin{cases} \mathbf{x}_i & i \in I_S, \\ \rho(\mathbf{x}_i \mid \mathbf{X}, S) & i \in I_{\bar{S}}. \end{cases} \quad (2)$$

The expectation in (2) is computed over the randomness from the reference mechanism ρ , which generates replacement coordinates $\{\hat{\mathbf{x}}_i\}_{i \in I_{\bar{S}}}$ for a given $\mathbf{X}$ and coalition S . With a deterministic

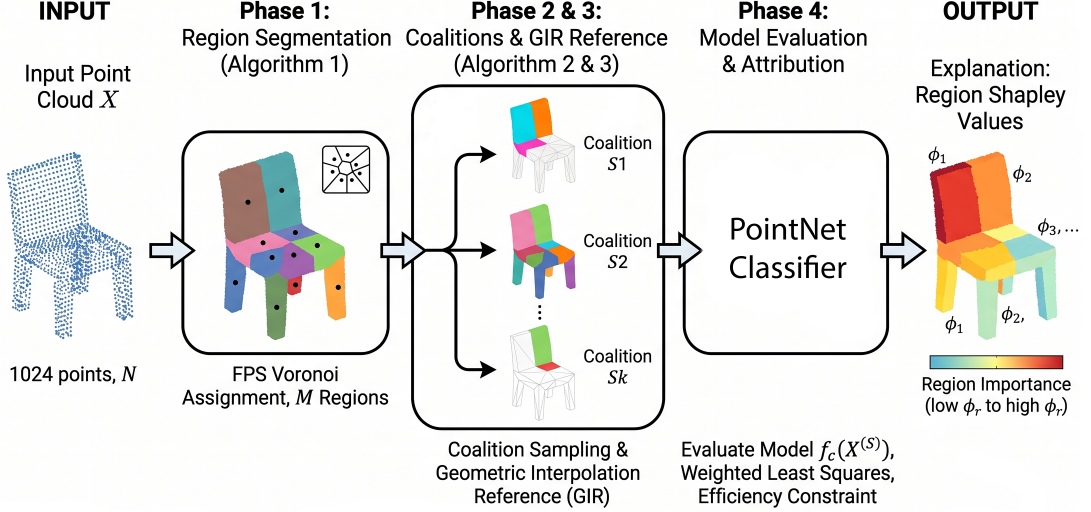


Fig. 1. R-SHAP pipeline. (1) An input point cloud is partitioned into M regions using FPS and a Voronoi assignment. (2,3) Region coalitions are sampled, and absent regions are filled by GIR. (4) PointNet [3] classifier evaluates each coalition. (5) Region-level Shapley values are recovered by Kernel SHAP weighted least squares.

reference like zero replacement, this becomes a single evaluation of f_c on the perturbed cloud. For stochastic GIR in Section 3.2 with noise variance $\tau^2 > 0$, it is a Monte Carlo average over Gaussian samples $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \tau^2 I_3)$. Fixing $\mathbf{X}$, $\mathcal{R}$, and ρ , the mapping $S \mapsto v_f(S)$ defines a set function on $\mathcal{M} = \{1, \dots, M\}$ with values in $\mathbb{R}$ and $v_f(\emptyset) \in \mathbb{R}$. This is the characteristic function of a cooperative game [20,14]. Substituting (2) into (1) gives the region Shapley value ϕ_{R_j} , which satisfies all cooperative game axioms [20,11]. The efficiency property yields the attribution $f_c(\mathbf{X}) = \phi_0 + \sum_{j=1}^M \phi_{R_j}$, where $\phi_0 = v_f(\emptyset)$.

3.2 Geometric Interpolation Reference (GIR)

Choosing ρ is the *absent-feature problem* [11,22]. Naive choices such as zero-replacement or random noise place inputs far off the data manifold of trained 3D classifiers and bias $v_f(S)$ [16,25]. We propose GIR, which replaces an absent point by a Gaussian-weighted average of its neighbors in the original point cloud,

$$\hat{\mathbf{x}}_i = \frac{\sum_{j \in \mathcal{N}_K(i) \cap I_S} w_{ij} \mathbf{x}_j}{\sum_{j \in \mathcal{N}_K(i) \cap I_S} w_{ij}} + \epsilon_i, \quad w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \quad (3)$$

where $\mathcal{N}_K(i)$ is the set of K -nearest neighbors of i in $\mathbf{X}$ and $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \tau^2 I_3)$ is small Monte Carlo noise. If no present neighbor is available, we fall back to the centroid of I_S . Because GIR depends

only on relative pairwise distances, it is rotation-equivariant. Concretely, $\rho(S, \mathbf{R}\mathbf{X}) = \mathbf{R}\rho(S, \mathbf{X})$ for every $\mathbf{R} \in SO(3)$.

3.3 Region Segmentation with FPS and Voronoi

A clustering scheme suitable for a region game must be deterministic, balanced, spatially coherent, and rotation-equivariant [16,15]. FPS serves as a region builder, as demonstrated by Shen *et al.* [21], and meets these requirements when used with a Voronoi assignment. After the point cloud is centered at its centroid, FPS selects M seeds that maximize the minimum pairwise distance. Each point is then assigned to its nearest seed, as described in Algorithm 1. The resulting cells are convex and approximately equal in size. They also remain identical under rotation because all distances are rotation-invariant.

Algorithm 1 Rotation-Equivariant Region Segmentation (FPS + Voronoi)

Require: Point cloud $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, number of regions M

Ensure: Region assignment $\{R_1, \dots, R_M\}$

```

1:  $\mathbf{c} \leftarrow \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$  ▷ Compute centroid
2:  $\mathbf{X}_c \leftarrow \{\mathbf{x}_i - \mathbf{c}\}_{i=1}^N$  ▷ Center the point cloud
3:  $\mathbf{s}_1 \leftarrow \arg \max_{\mathbf{x} \in \mathbf{X}_c} \|\mathbf{x}\|$  ▷ First seed farthest from centroid
4: for  $j = 2$  to  $M$  do
5:    $\mathbf{s}_j \leftarrow \arg \max_{\mathbf{x} \in \mathbf{X}_c} \min_{k < j} \|\mathbf{x} - \mathbf{s}_k\|$  ▷ Farthest from existing seeds
6: end for
7: for  $i = 1$  to  $N$  do
8:    $r(i) \leftarrow \arg \min_j \|\mathbf{x}_i - \mathbf{s}_j\|$  ▷ Voronoi assignment
9: end for
10: return  $R_j = \{i : r(i) = j\}$  for  $j = 1, \dots, M$ 

```

3.4 Kernel R-SHAP

For M players the coalition space contains 2^M subsets. Even with $M = 8$ this is only 256 coalitions, but a stable estimate can be obtained from far fewer evaluations. We follow the Kernel SHAP construction of Lundberg and Lee [11] and recover ϕ as the solution of the weighted least squares problem

$$\hat{\phi} = \arg \min_{\phi} \sum_S \pi(S) \left(v_f(S) - \phi_0 - \sum_{j \in S} \phi_j \right)^2, \quad \pi(S) = \frac{M-1}{\binom{M}{|S|} |S| (M-|S|)}, \quad (4)$$

subject to the efficiency constraint $\sum_j \phi_j = f_c(\mathbf{X}) - v_f(\emptyset)$. The kernel $\pi(S)$ is the unique weighting that recovers (1) in expectation [11]. Algorithm 2 details the sampling routine, taking the regions $\mathcal{R}$ produced by Algorithm 1 and a fixed reference mechanism ρ as inputs and returning the region attributions. Coalition evaluations dominate the runtime and are easily batched on a GPU.

Algorithm 2 Kernel R-SHAP. Sampling-Based Estimation of Region Shapley Values**Require:** Point cloud $\mathbf{X}$, classifier f , class c , regions $\mathcal{R} = \{R_1, \dots, R_M\}$, samples K , reference ρ **Ensure:** Region Shapley values $\hat{\phi} \in \mathbb{R}^M$

- 1: Initialise $\mathbf{Z} \in \{0, 1\}^{K \times M}$, $\mathbf{y} \in \mathbb{R}^K$, $\mathbf{w} \in \mathbb{R}^K$
- 2: **for** $k = 1$ to K **do**
- 3: Draw a coalition size $|S|$ stratified by the kernel π
- 4: Sample S uniformly from coalitions of that size
- 5: $\mathbf{z}_k \leftarrow \mathbf{1}_S$ ▷ Binary coalition mask
- 6: $\mathbf{X}^{(S)} \leftarrow \rho(\mathbf{X}, S)$ ▷ Apply reference mechanism
- 7: $y_k \leftarrow f_c(\mathbf{X}^{(S)})$ ▷ Forward pass
- 8: $w_k \leftarrow \pi(S)$ ▷ Shapley kernel weight
- 9: **end for**
- 10: Solve $\hat{\phi} \leftarrow \arg \min_{\phi} \sum_k w_k (y_k - \phi_0 - \mathbf{z}_k^\top \phi)^2$
- 11: subject to $\sum_j \phi_j = f_c(\mathbf{X}) - v_f(\emptyset)$
- 12: **return** $\hat{\phi}$

3.5 R-SHAP Pipeline

The four stages illustrated in Fig. 1 are integrated within Algorithm 3. The first phase segments the point cloud into regions using Algorithm 1. The second phase samples coalitions of regions based on the Shapley kernel. The third phase applies GIR in order to fill the absent regions of each coalition. The fourth phase evaluates the classifier on each perturbed point cloud and uses the results as input for the kernel regression in Algorithm 2. Segmentation is performed once for each point cloud, which allows the computation to be shared across all coalition evaluations.

Algorithm 3 R-SHAP. Region-Based Shapley Explanations for Point Clouds**Require:** Point cloud $\mathbf{X} \in \mathbb{R}^{N \times 3}$, classifier f , target class c **Require:** Hyperparameters regions M , samples K , kernel width σ , noise τ **Ensure:** Region Shapley values $\hat{\phi} \in \mathbb{R}^M$

- 1: **Phase 1. Region Segmentation**
- 2: $\mathcal{R} \leftarrow \text{FPS_Voronoi}(\mathbf{X}, M)$ ▷ Algorithm 1
- 3: **Phase 2. Coalition Sampling**
- 4: Draw K coalitions $\{S_k\}_{k=1}^K$ stratified by $|S|$ with kernel $\pi(S)$
- 5: **Phase 3. Reference Mechanism**
- 6: **for** each coalition S_k **do**
- 7: $\mathbf{X}^{(S_k)} \leftarrow \text{GIR}(\mathbf{X}, S_k, \sigma, \tau)$ ▷ Equation (3)
- 8: **end for**
- 9: **Phase 4. Model Evaluation and Regression**
- 10: **for** each coalition S_k **do**
- 11: $y_k \leftarrow f_c(\mathbf{X}^{(S_k)})$ ▷ Forward pass through PointNet
- 12: **end for**
- 13: $\hat{\phi} \leftarrow \text{KernelRSHAP}(\mathbf{X}, f, c, \mathcal{R}, K, \rho)$ ▷ Algorithm 2
- 14: **return** $\hat{\phi}$

3.6 Properties and Faithfulness Evaluation

R-SHAP inherits the four Shapley axioms, which are efficiency, symmetry, null-player, and linearity, directly from (1). When these axioms are combined with the rotation-equivariance of GIR and FPS + Voronoi assignment, the explanation remains rotation-equivariant provided that f is rotation-invariant. Every part of the value function transforms consistently under any $\mathbf{R} \in SO(3)$, as shown by the fact that $\phi_{R_j}(v_f^{\mathbf{R}\mathbf{X}}) = \phi_{R_j}(v_f^{\mathbf{X}})$. This equivariance is approximate, since PointNet’s T-Net is only approximately invariant.

To assess the quality of explanations, we adopt the deletion and insertion methodology that is standard in XAI evaluation [17,15]. We rank regions in descending order according to ϕ_{R_j} . The deletion curve $D(k)$ measures the prediction after sequentially removing the top k regions. The insertion curve $I(k)$ captures the prediction obtained by gradually re-inserting these regions into the empty baseline. Their normalized areas

$$\text{AUDC} = \frac{1}{M\Delta_f} \sum_{k=0}^{M-1} D(k), \quad \text{AUCI} = \frac{1}{M\Delta_f} \sum_{k=1}^M I(k), \quad \mathcal{F} = \text{AUCI} - \text{AUDC}, \quad (5)$$

with $\Delta_f = |f_c(\mathbf{X}) - f_c(\mathbf{X}^{(0)})|$, take values in $[-1, 1]$. A faithful explanation has the prediction collapse quickly under deletion and recover quickly under insertion, yielding $\mathcal{F}$ close to 1. In our experiments, the faithfulness curves in (5) use the softmax probability so that AUDC, AUCI, and $\mathcal{F}$ remain within $[-1, 1]$, whereas the targeted ablation in Section 4.4 uses raw logits.

4 Results and Discussion

We implement R-SHAP on the ModelNet10 benchmark [26] using a PointNet classifier [3] that is trained from scratch on the official train and test splits with 1024 points per point cloud. The network follows the standard PointNet architecture, which includes two T-Nets, five 1×1 convolutions with widths 64, 64, 64, 128, 1024, a global max-pool, and three fully connected layers. The model is optimized with Adam (learning rate 10^{-4} , weight decay 10^{-4}) and trained under a cross-entropy loss for 20 epochs using a step-decay schedule. The training loss decreases steadily from 1.54 at the first epoch to 0.32 at the twentieth epoch. Test accuracy improves from 51.4% to a maximum of 87.89%, with a final epoch accuracy of 87.11%. All R-SHAP experiments are conducted with $M = 8$ regions, $K = 1000$ Kernel SHAP samples, GIR using $K_{\text{nn}} = 20$, and a logit value space.

4.1 Manifold Preservation under Geometric Interpolation

A region reference is effective only if it remains close to the manifold of point clouds on which the classifier was trained. We evaluate GIR by randomly removing half of the points in a point cloud and then tasking each candidate reference with filling in the missing portion. On the analytic unit-sphere manifold, GIR maintains the interpolated points within an average of 0.028 from the true sphere, with a maximum deviation of 0.078. In contrast, zero-replacement causes all absent points to collapse to the origin, resulting in a mean distance of 1.000, while Gaussian noise disperses these points away from the surface and leads to a mean of 0.835. Fig. 2 presents visualizations for all three reference strategies.

The same behavior is observed with real data. When half of a bathtub is removed and then reconstructed using GIR, the Chamfer distance to the original surface is 0.0298 as shown in Fig. 3.

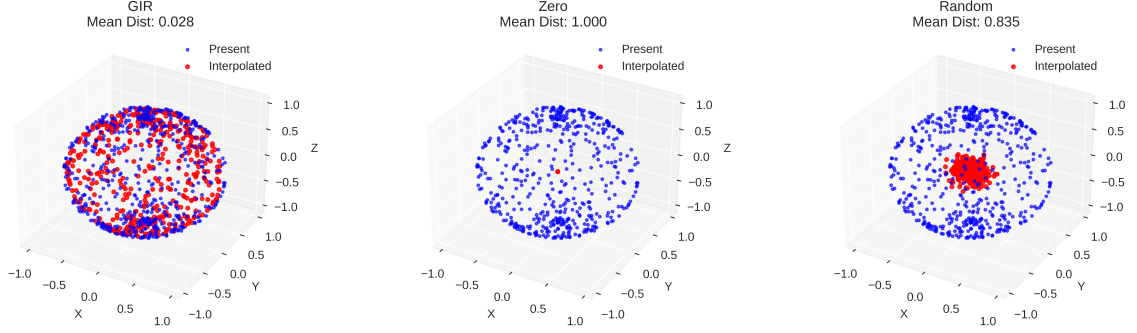


Fig. 2. GIR validation on the analytic unit-sphere manifold. Half of the points are removed and re-imputed by GIR, zero-replacement, or Gaussian noise. GIR (left) keeps interpolated points (red) on the sphere, zero replacement (centre) maps them all to the origin, and random noise (right) scatters them off the surface.

Table 1. Chamfer distance between GIR-completed and original clouds when 50% of points are removed, reported per class. Lower is better.

Class	Chamfer Distance	Class	Chamfer Distance
bathtub	0.0298	monitor	0.0196
bed	0.0189	night stand	0.0167
chair	0.0153	sofa	0.0202
desk	0.0222	table	0.0197
dresser	0.0226	toilet	0.0132
Mean Chamfer distance = 0.0198			

The imputed points, which are depicted in red, align with the bathtub rim instead of collapsing toward the origin. This experiment was repeated with one sample from each class, and the results are presented in Table 1. The mean Chamfer distance is 0.0198, with values ranging from 0.0132 for the toilet to 0.0298 for the bathtub. Smaller and more homogeneous shapes, such as toilets, chairs, and night stands, are easier to compare to more topologically complex shapes, such as bathtubs and dressers. In each case, GIR produces a manifold deviation that is an order of magnitude smaller than the deviation produced by either the zero reference or the random reference.

4.2 Faithfulness via Deletion and Insertion Curves

We evaluate faithfulness using the deletion and insertion areas defined by (5). For each of the ten classes, we calculate AUDC, AUIC, and $\mathcal{F} = \text{AUIC} - \text{AUDC}$, and we compare these results with a random-ranking baseline. The deletion and insertion curves are presented in Fig. 4, and the class-specific values are provided in Table 2. R-SHAP achieves a mean faithfulness of $\mathcal{F} = 0.310 \pm 0.255$ across all ten classes, compared to $\mathcal{F}_{\text{rand}} = 0.033$ for random ordering. The greatest improvements are observed for *night stand* ($\mathcal{F} = +0.717$), *table* (+0.710), and *bathtub* (+0.509), where only a few regions carry most of the predictive evidence. The two classes in which R-SHAP performs below the random baseline, namely *bed* and *toilet*, are precisely those with very low prediction confidence ($p_c = 0.287$ and 0.482). When the original prediction is already weak, the deletion and insertion

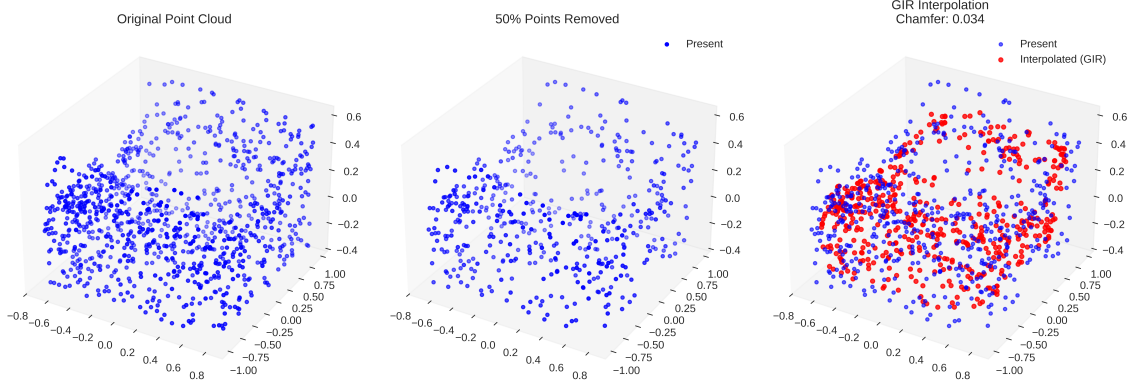


Fig. 3. GIR validation on the bathtub class. The original point cloud (left), the point cloud after removing 50% of the points (middle), and the GIR-completed point cloud (right). The imputed points (red) adopt the local geometry of nearby points, resulting in a Chamfer distance of 0.034 to the original surface.

areas exhibit a low signal-to-noise ratio, and thus the random ranking can sometimes match R-SHAP by chance.

The two cases in which R-SHAP performs below the random baseline, namely *bed* ($\mathcal{F} = +0.033$ compared to $+0.263$ random) and *toilet* ($+0.112$ compared to $+0.475$ random), correspond to the lowest prediction confidence values in the test set: $p_c = 0.287$ for *bed* and $p_c = 0.482$ for *toilet*. When confidence is already below 0.5, the total prediction change $\Delta_f = |f_c(\mathbf{X}) - f_c(\mathbf{X}^{(0)})|$ used to normalise the curves in (5) remains small, limiting the extent to which the deletion curve can decrease before reaching saturation. For *bed*, the empty baseline prediction actually exceeds the original prediction ($f_c(\mathbf{X}^{(0)}) = 0.970$ compared to $f_c(\mathbf{X}) = 0.287$), resulting in the insertion curve starting near its maximum and rendering any region ordering apparently faithful.

Under these circumstances, a random ranking may match or surpass the Shapley ranking by chance, as reported alongside the faithfulness score in Table 2. This pattern aligns with the findings of Mulawade et al. [16], who observed that perturbation-based explanations tend to track the classifier rather than the input. Consequently, *bed* and *toilet* are identified as classes that require a more robust classifier rather than an alternative explanation method. In contrast, the *dresser* case demonstrates that R-SHAP outperforms random ordering ($\mathcal{F} = +0.074$ compared to -0.057 random), although both scores remain close to zero due to the roughly equal informativeness of the dresser regions.

4.3 Critical-Point Distribution in PointNet

PointNet aggregates features using a global max-pool, which means its prediction depends on a small set of *critical points* that achieve the channel-wise maxima [3]. For one sample from each class, we examine the distribution of these critical points across the eight FPS + Voronoi regions. We report the entropy-based concentration $1 - H(p)/\log M$, where a value of 0 indicates a uniform distribution and a value of 1 indicates that all critical points are located in a single region. Across the ten classes, we observe an average of 269 unique critical points selected across the 1024 feature channels and a mean concentration of only 0.063. The per-class values are as follows: bathtub 0.045,

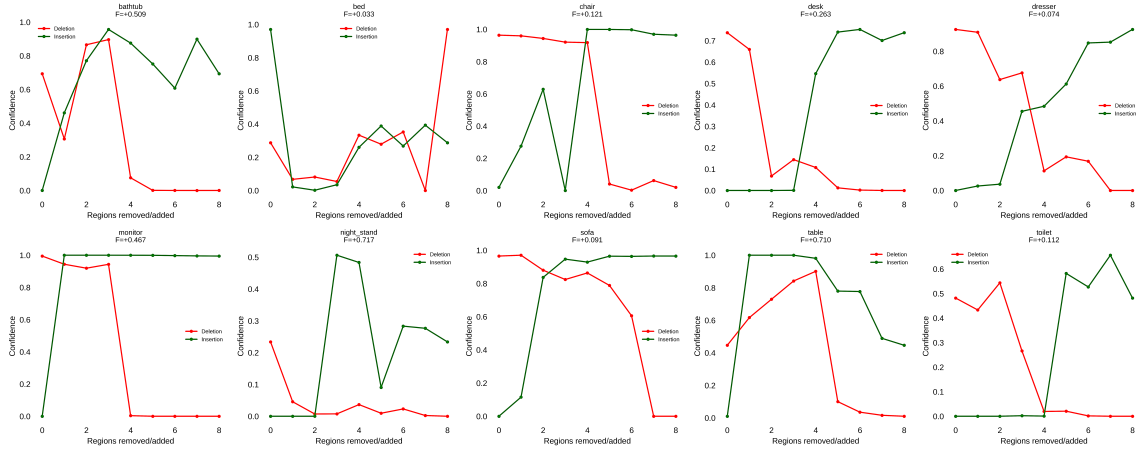


Fig. 4. Deletion and Insertion Curves for Each Class

Table 2. Per-class faithfulness for R-SHAP. $\mathcal{F} > 0$ means R-SHAP outperforms the random ranking.

Class	AUDC	AUIC	$\mathcal{F}$	Random	Class	AUDC	AUIC	$\mathcal{F}$	Random
bathtub	0.455	0.964	+0.509	-0.390	monitor	0.425	0.892	+0.467	-0.297
bed	0.395	0.427	+0.033	+0.263	night stand	0.174	0.891	+0.717	-0.255
chair	0.569	0.689	+0.121	+0.015	sofa	0.679	0.770	+0.091	-0.152
desk	0.261	0.524	+0.263	+0.138	table	0.941	1.650	+0.710	+0.591
dresser	0.435	0.509	+0.074	-0.057	toilet	0.407	0.519	+0.112	+0.475
Mean $\mathcal{F} = 0.310$, mean random $\mathcal{F} = 0.033$									

bed 0.044, chair 0.085, desk 0.091, dresser 0.057, monitor 0.076, night stand 0.025, sofa 0.084, table 0.076, and toilet 0.047.

This observation provides a direct explanation from the model’s perspective regarding why R-SHAP attributions for certain classes, such as *sofa* and *bed*, are smaller and flatter than for others. PointNet does not depend disproportionately on any single spatial region because its max-pool aggregates supporting evidence from points that are distributed throughout the object. As a result, R-SHAP accurately captures this property and does not introduce artificial contrast that is absent from the model’s internal reasoning.

4.4 Targeted Region Ablation

We complete the analysis by removing the region with the largest Shapley value, the region with the smallest value, and a randomly selected region. We then measure the resulting change in the predicted-class logit, which is the same value space used to compute the Shapley estimates. Fig. 6 and the per-class results in Table 3 indicate that removing the most important region leads to a mean logit drop of 0.517. In contrast, removing a random region results in a smaller drop of 0.211, which is 2.45 times less. Removing the least important region, on average, produces a mean logit drop of -0.179 (i.e., confidence increases by 0.179). This sign reversal is consistent with theoretical

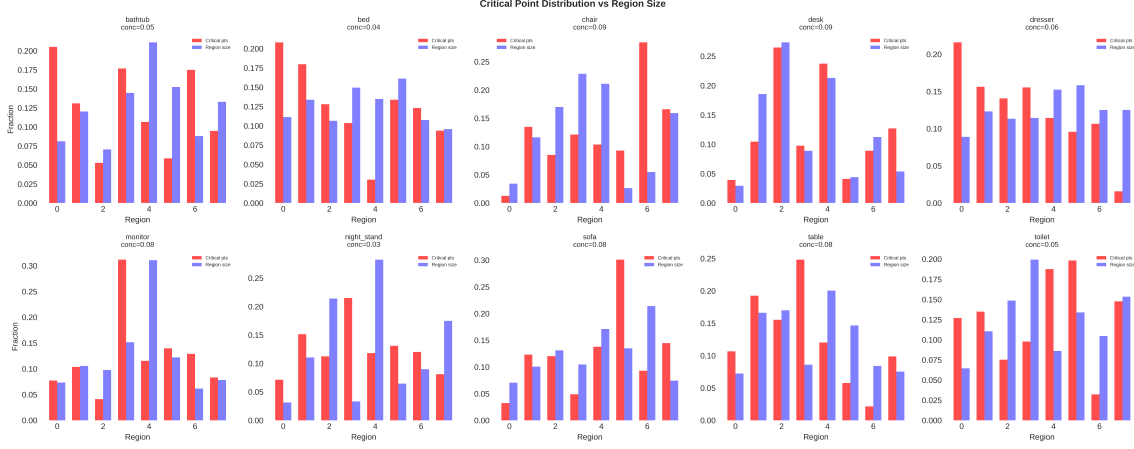


Fig. 5. Distribution of PointNet critical points (red) against region size (blue) per class. Both distributions are nearly identical, giving a mean concentration of 0.063. Critical points are spread across all regions, which is consistent with the moderately differentiated R-SHAP attributions reported in Table 4.

Table 3. Logit drop after removing the most and least important and a random region (averaged over three random regions). Larger “most” is better, and “least” is expected to be small or negative.

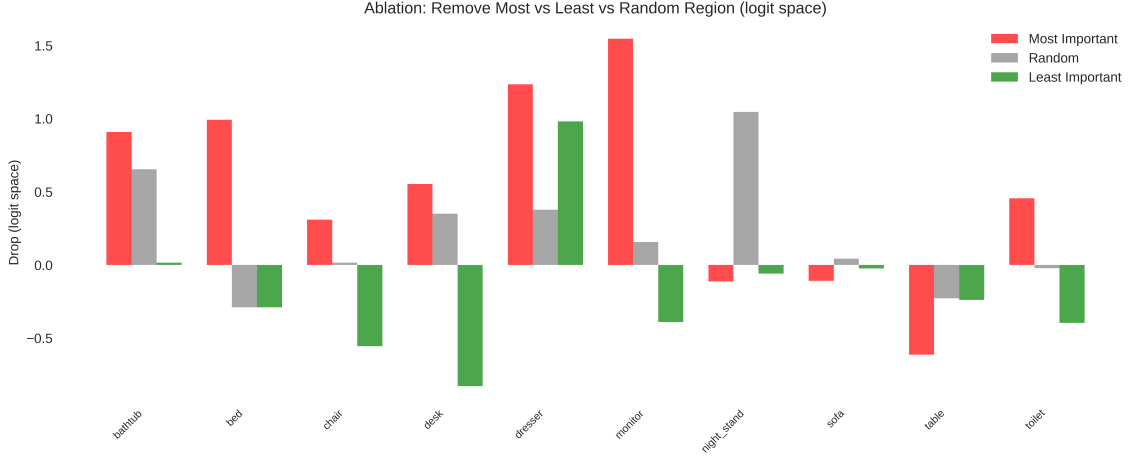
Class	Most	Least	Random	Class	Most	Least	Random
bathtub	+0.909	+0.015	+0.654	monitor	+1.546	−0.390	+0.157
bed	+0.992	−0.290	−0.290	night stand	−0.113	−0.059	+1.045
chair	+0.309	−0.555	+0.016	sofa	−0.108	−0.024	+0.043
desk	+0.553	−0.828	+0.350	table	−0.613	−0.240	−0.228
dresser	+1.235	+0.981	+0.378	toilet	+0.456	−0.396	−0.021
Mean: most 0.517, least −0.179, random 0.211; Ratio most/random = 2.45×							

expectations. Regions with the largest negative Shapley value provide evidence against the predicted class, so removing them benefits the model’s confidence.

The analysis of the weak cases (*sofa* and *table*) is informative. Both classes exhibit the smallest absolute Shapley range in Table 4 and display the most distributed critical points in Fig. 5. As a result, the difference between the most important region and a random region is small by design. This observation reflects a characteristic of PointNet for these classes rather than a limitation of R-SHAP. For the other eight classes, the ablation of the most important region consistently produces a greater drop than the ablation of a random region, with the largest effects observed for *monitor* (+1.55), *dresser* (+1.24), and *bed* (+0.99).

4.5 Per-Class Qualitative Explanations and Attribution Statistics

Fig. 7 presents the eight FPS + Voronoi regions and the associated Shapley bar chart for representative *chair* and *table* samples. The region partitions are spatially coherent and approximately balanced, as anticipated in Section 3.3. The distributions of importance correspond to the part

**Fig. 6.** Targeted Region Ablation**Table 4.** R-SHAP Attribution Statistics per Class.

Class	Min	Max	Mean	Standard Deviation	Range	Positive Ratio
bathtub	-1.0062	+2.0270	+0.6717	0.9976	3.0332	0.750
bed	-1.5435	-0.5328	-0.9279	0.2994	1.0107	0.000
chair	-2.1232	+3.5437	+0.0429	1.6000	5.6669	0.375
desk	-1.1065	+4.5063	+1.2376	2.2582	5.6128	0.375
dresser	-1.3143	+4.2127	+2.3513	1.5645	5.5271	0.875
monitor	-1.7138	+7.4646	+1.7522	3.2931	9.1784	0.500
night stand	-0.2338	+3.7534	+1.6198	1.3691	3.9872	0.875
sofa	+0.3816	+1.8485	+1.1243	0.5159	1.4670	1.000
table	-4.1475	+2.7959	+0.0658	2.4340	6.9433	0.500
toilet	+0.0081	+3.3841	+0.9151	1.0930	3.3760	1.000

structure of the underlying object. In the case of the chair, the seat and backrest region is dominant with $\phi = +3.5437$. For the table, different leg and edge regions carry positive and negative attributions. This pattern reflects the diagnostic challenges associated with the *table* class, which were also noted in Section 4.4.

Per-class summary statistics for all ten classes, as shown in Table 4, reveal substantial variation in the range of importance values. The range spans from 1.01 for *bed* to 9.18 for *monitor*. The positive ratio, defined as the fraction of regions with $\phi_{R_j} > 0$, reaches 1.0 for both *sofa* and *toilet*. In these two classes, every region contributes positively to the prediction. This observation aligns with the low concentration of critical points discussed in Section 4.3 and the moderate ablation drops described in Section 4.4. In contrast, *table* and *chair* display the largest negative attributions, namely -4.15 and -2.12 . Whereas *chair* and *desk* have the lowest positive ratio of 0.375. These results are consistent with the more challenging ablation behavior observed for these classes.

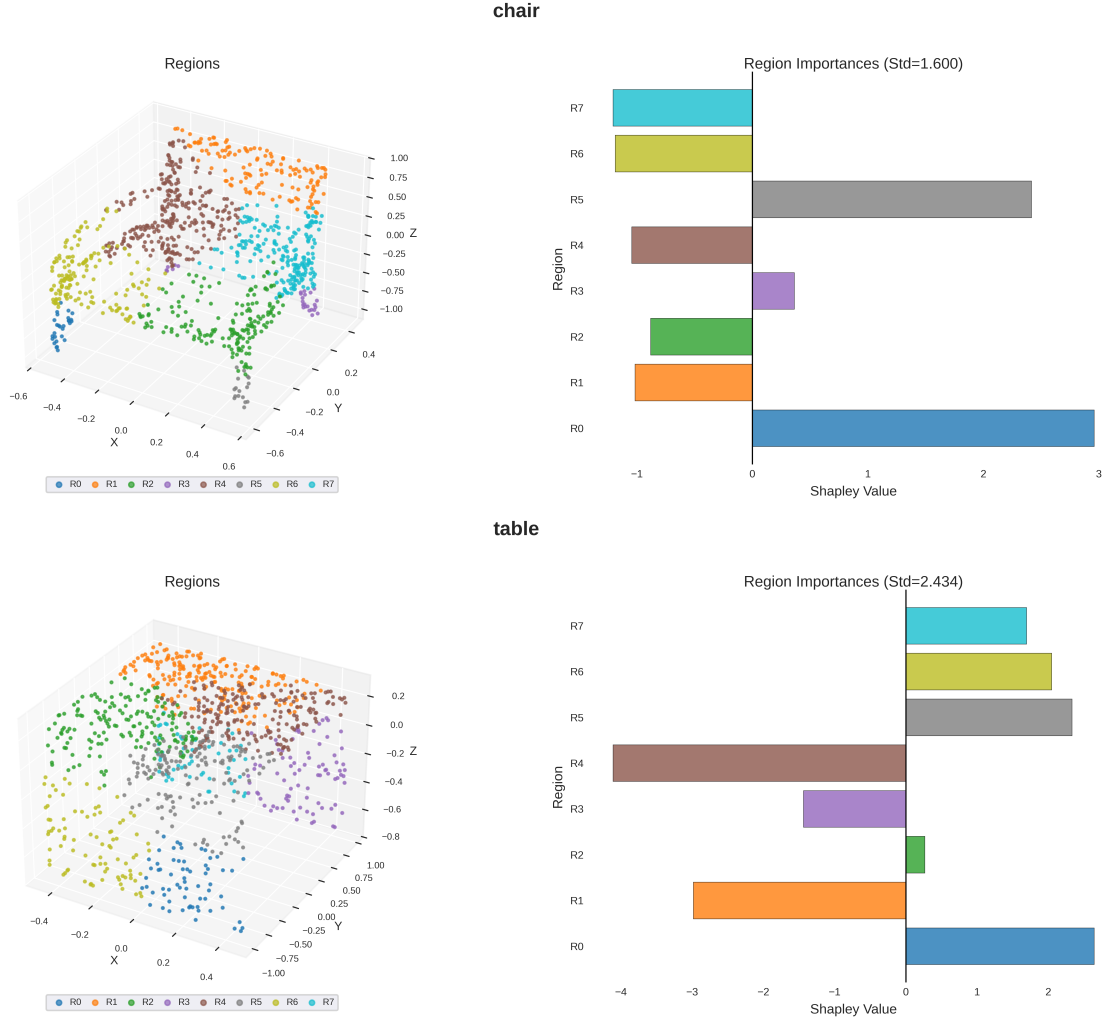


Fig. 7. Per-class qualitative R-SHAP explanations for *chair* and *table*. Each panel shows the eight FPS + Voronoi regions on the left and the corresponding Shapley values on the right. Regions with positive Shapley values support the predicted class, and negative values are evidence against it.

5 Conclusion

R-SHAP provides explanations for 3D point clouds by reformulating Shapley attribution as a cooperative game over regions. The method uses rotation-equivariant segmentation, manifold-faithful interpolation, and Kernel SHAP regression. On ModelNet10 with PointNet, R-SHAP achieves a mean $\mathcal{F}$ of 0.310 and targeted ablation confirms attribution faithfulness. Results show GIR maintains perturbations close to the data manifold and PointNet critical points are distributed across

regions, clarifying the shape of R-SHAP attributions. The pipeline is model-agnostic and adaptable to other architectures and datasets.

Declarations

Conflict of Interests. The authors declare that they have no conflict of interest.

Data Availability. In this study, a public dataset is used. The public dataset is available at <https://modelnet.cs.princeton.edu>.

Code Availability. Our code is available at <https://github.com/mshoaibsarwar/region-based-explanations>.

AI Statement. The authors used Grammarly to improve the work’s academic writing style and structure. Objects in Fig. 1 were generated with Gemini, and the prompt will be provided upon request. The authors carefully reviewed all suggestions and edited the content as needed.

References

1. Arnold, N.I., Angelov, P., Atkinson, P.M.: An improved explainable point cloud classifier (xpcc). *IEEE Transactions on Artificial Intelligence* **4**(1), 71–80 (2023). <https://doi.org/10.1109/TAI.2022.3150647>
2. Atik, M.E., Duran, Z., Seker, D.Z.: Explainable artificial intelligence for machine learning-based photogrammetric point cloud classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **17**, 5834–5846 (2024). <https://doi.org/10.1109/JSTARS.2024.3370159>
3. Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 77–85 (2017). <https://doi.org/10.1109/CVPR.2017.16>
4. Günen, M.A.: Explainable artificial intelligence for rock discontinuity detection from point cloud with ensemble methods. *Journal of Rock Mechanics and Geotechnical Engineering* **17**(12), 7590–7611 (2025). <https://doi.org/10.1016/j.jrmge.2025.01.052>
5. Gupta, A., Watson, S., Yin, H.: 3d point cloud feature explanations using gradient-based methods. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2020). <https://doi.org/10.1109/IJCNN48605.2020.9206688>
6. Harinck, M.F., Sarwar, M.S., Ton, B., Ahmed, F.: SPAX: A shapley-based point attribution explanation for interpreting 3D point cloud classification. In: Marcelloni, F., Madani, K., van Stein, N., Filipe, J. (eds.) *Computational Intelligence*. pp. 414–433. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-15638-9_24
7. Holzinger, A., Lukač, N., Rozajac, D., Johnston, E., Kocic, V., Hoerl, B., Gollob, C., Nothdurft, A., Stampfer, K., Schweng, S., Ser, J.D.: Enhancing trust in automated 3D point cloud data interpretation through explainable counterfactuals. *Information Fusion* **119**, 103032 (2025). <https://doi.org/10.1016/j.inffus.2025.103032>
8. Kalasampath, K., KN, S., Sajeev, S., Kuppa, S., Ajay, K., Angulakshmi, M.: A literature review on applications of explainable artificial intelligence (xai). *IEEE Access* **PP**, 1–1 (2025). <https://doi.org/10.1109/ACCESS.2025.3546681>
9. Kuroki, M., Yamasaki, T.: Explaining 3d object detection through shapley value-based attribution map. In: 2024 IEEE International Conference on Image Processing (ICIP). pp. 507–513 (2024). <https://doi.org/10.1109/ICIP51287.2024.10647574>
10. Levi, M.Y., Gilboa, G.: Fast and simple explainability for point cloud networks (2024), <https://arxiv.org/abs/2403.07706>

11. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 30, pp. 4765–4774. Curran Associates, Inc. (2017)
12. Matrone, F., Paolanti, M., Felicetti, A., Martini, M., Pierdicca, R.: Bubblex: An explainable deep learning framework for point-cloud classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **15**, 6571–6587 (2022). <https://doi.org/10.1109/JSTARS.2022.3195200>
13. Matrone, F., Paolanti, M., Frontoni, E., Pierdicca, R.: Enhancing explainability of deep learning models for point cloud analysis: a focus on semantic segmentation. *International Journal of Digital Earth* **17**(1), 2390457 (2024). <https://doi.org/10.1080/17538947.2024.2390457>
14. Molnar, C.: *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Lean Publishing (2019)
15. Mulawade, R.N., Garth, C., Wiebel, A.: Explainable artificial intelligence (XAI) for methods working on point cloud data. a survey. *IEEE Access* **12**, 146830–146851 (2024). <https://doi.org/10.1109/ACCESS.2024.3472872>
16. Mulawade, R.N., Garth, C., Wiebel, A.: Xai for point cloud data using perturbations based on meaningful segmentation. *IEEE Access* **13**, 140175–140186 (2025). <https://doi.org/10.1109/access.2025.3597094>
17. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM computing surveys* **55**(13s) (2023). <https://doi.org/10.1145/3583558>
18. Saranti, A., Pfeifer, B., Gollob, C., Stampfer, K., Holzinger, A.: From 3d point-cloud data to explainable geometric deep learning: State-of-the-art and future challenges. *WIREs Data Mining and Knowledge Discovery* **14**(6), e1554 (2024). <https://doi.org/10.1002/widm.1554>
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Oct 2017)
20. Shapley, L.S.: A value for n-person games. *Contributions to the Theory of Games II* pp. 307–317 (1953). <https://doi.org/10.1515/9781400881970-018>
21. Shen, W., Ren, Q., Liu, D., Zhang, Q.: Interpreting representation quality of DNNs for 3D point cloud processing. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21*, Curran Associates Inc., Red Hook, NY, USA (2021). <https://doi.org/10.5555/3540261.3540939>
22. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* **41**, 647–665 (2014). <https://doi.org/10.1007/s10115-013-0679-x>
23. Taghanaki, S.A., Hassani, K., Jayaraman, P.K., Khasahmadi, A.H., Custis, T.: Pointmask: Towards interpretable and bias-resilient point cloud processing (2020), <https://arxiv.org/abs/2007.04525>
24. Tan, H., Kotthaus, H.: Surrogate model-based explainability methods for point cloud nns. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 2239–2248 (January 2022)
25. Tian, Z., Li, W., Deng, C., Hu, J., Wang, Y., Liu, W.: Desensitizing for improving corruption robustness in point cloud classification through adversarial training. *Pattern Recognition* **169**, 111862 (2026). <https://doi.org/10.1016/j.patcog.2025.111862>
26. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D ShapeNets: A deep representation for volumetric shapes (2015). <https://doi.org/10.1109/CVPR.2015.7298801>, <https://modelnet.cs.princeton.edu/>
27. Xia, H., Li, X., Pang, J., Liu, J., Ren, K., Xiong, L.: P-shapley: Shapley values on probabilistic classifiers. *Proc. VLDB Endow.* **17**(7), 1737–1750 (2024). <https://doi.org/10.14778/3654621.3654638>