Surrogate Model-Based Explainability Methods for Point Cloud NNs

Hanxiao Tan Helena Kotthaus
AI Group, TU Dortmund
{hanxiao.tan,helena.kotthaus}@tu-dortmund.de

Abstract

In the field of autonomous driving and robotics, point clouds are showing their excellent real-time performance as raw data from most of the mainstream 3D sensors. Therefore, point cloud neural networks have become a popular research direction in recent years. So far, however, there has been little discussion about the explainability of deep neural networks for point clouds. In this paper, we propose a point cloud-applicable explainability approach based on a local surrogate model-based method to show which components contribute to the classification. Moreover, we propose quantitative fidelity validations for generated explanations that enhance the persuasive power of explainability and compare the plausibility of different existing point cloud-applicable explainability methods. Our new explainability approach provides a fairly accurate, more semantically coherent and widely applicable explanation for point cloud classification tasks. Our code is available at `https://github.com/Explain3D/LIME-3D`

1. INTRODUCTION

Deep neural networks (DNNs) have risen on the stage of machine learning in recent years with prominent accuracy and omnipotent end-to-end learning capability. Especially in the field of computer vision, complex structured neural networks show better image recognition performance than humans. Despite their great success in industry, DNNs suffer from the trade-off between performance and explainability [7] due to the nonlinear model architectures. With an increasing demand for the credibility of decision-making, studies of explainability for black-box models have received considerable critical attention. Existing research recognizes that models with high explainability play a crucial role in gaining user confidence, exposing potential biases in training data, and improving model robustness [5].

Several studies have found intimate connections between explainability and the safety of human life in safety-critical areas, e.g., in medicine [10, 18, 2] or autonomous driving [25, 9]. In the medical domain, decisions made by black-box models are unreliable and thus unacceptable [28]. The same dilemma occurs in the field of autonomous driving, where algorithms that control vehicles with low transparency not only lead to legal problems but also pose a potential social threat [13]. Therefore, both customers and companies benefit from the research in explainable machine learning. Recent studies have proposed several explainability approaches for explaining complex machine learning models, among which the most popular methods are gradient-based [20, 38, 41, 4, 29, 34] and local surrogate model-based [26, 15].

Point cloud data, as the raw data of most mainstream sensors, has a significant advantage in real-time scenarios compared to other 3D data formats and therefore has become a popular research direction in recent years. Point clouds exhibit higher structural complexity than 2D images. For instance, convolution kernels are easily applied to images due to their regularity, but they are not directly applicable to point clouds. Due to the lack of adjacency of the point cloud data, neighboring points in the point cloud matrix have a high probability of being irrelevant to the 3D spatial adjacency, which leads to the invalidation of the traditional convolution kernel. [23, 24, 22] bring up solutions for point feature extraction and make point clouds suitable for convolutional neural networks. However, in contrast to the field of 2D image processing with a large number of explainability studies [2, 37, 8, 35], most point cloud-compatible DNNs currently remain black-boxes due to the paucity of research investigating their working principle [12]. This indicates an indispensable need for the explainability research on DNNs dealing with point cloud data to ensure transparency of decisions made by robots and autonomous vehicles.

As for the reliability examination of explainability methods, to date, there is no acknowledged evaluation criterion. Most of the previous work validates the explanation results subjectively based on human interpretation, which easily leads to bias in the evaluation of explainability approaches. Therefore, quantitative evaluations are increasingly recognized as an essential requirement in the explainable machine learning domain.

This work proposes a point cloud-applicable *local sur-*

rogate model-based approach [26, 15] investigating the explainability and reliability of point cloud neural networks. With the help of explanations, humans gain a better awareness of the underlying reasons for misclassification cases. Besides, we quantify the plausibility of the explanations for point cloud data through fidelity and accuracy verification methods instead of a subjective approach based on human interpretation. Our contribution is primarily summarized as follows:

- We propose a local surrogate model-based explainability approach for point cloud DNNs based on LIME [26], which is more widely applicable than gradient-based methods [12].
- We provide two quantitative evaluations for 3D explanations: fidelity metrics and cluster flipping, which are applied to validate the fidelity and plausibility of surrogate model-based and all 3D explainability approaches, respectively.
- We present quantitative comparisons of our proposed method with existing approaches for point cloud data using the proposed evaluation approach. Besides, we demonstrate an interesting viewpoint on the misclassified cases through our proposed approach.

The overall structure of this paper takes the form of five sections: In section 2, we introduce the outline of existing explainability methods and 3D neural networks, and the possibility of validating explainability approaches. Section 3 sets out details of our explainability approaches and the corresponding verification metrics. In section 4, we present the qualitative and quantitative results of our proposed methods. In section 5, we conclude a brief summary and suggest future research directions.

2. RELATED WORK

This section reviews the current widely used explainability approaches, summarizes the classical point cloud neural network, presents existing explainability methods for point cloud DNNs, and identifies the current possibilities for verifying the explainability approaches.

Explainability approaches: Most of the current research on explainability pays particular attention to image classification tasks. Popular methods for explaining DNNs are gradient-based and local surrogate model-based.

Gradient-based approaches observe the process of gradient descent during forward passes. Therefore they are only applicable to differentiable models such as neural networks. Saliency Maps [30] is the pioneer who attempts to explain DNNs by computing the partial derivative of each pixel of the image as its attribution. However, vanilla gradients suffer from saturation [33] and discontinuity [31]. Integrated Gradient [34], Layer-wise Relevance Propagation (LRP) [4] and DeepLIFT [29] solve the saturated gradient problem by estimating the global importance of each pixel [14]. On the other hand, SmoothGrad [31] relieves the discontinuity issue by smoothing the discontinuous gradient with a Gaussian kernel that randomly samples the input neighbors and computes their average gradients. Guided Backpropagation [32] provides sharper gradient maps by removing gradients that have negative attributions to the prediction.

Another series of approaches that utilize the gradients is activation maximization [20]. Instead of explaining individual instances (local explanation), it attempts to discover the ideal input distribution of a given class (global explanation) by optimizing the gradients of the inputs while freezing all parameters of the networks.

Local surrogate model-based methods such as LIME [26] and KernelSHAP [15] aim to track the decision boundary around the selected instance by perturbing input instances and feeding them into surrogate linear models that approximate the performance of the original one but are more explainable due to their simplicity. Our work aims to provide a point cloud-applicable explainability approach based on local surrogate model-based methods as they are applicable to arbitrary machine learning models, which provides users with more choices.

3D convolutional neural networks: Recent developments in the field of robotics and autonomous driving have led to an increasing interest in 3D deep learning. Processing raw point cloud data efficiently plays an important role in designing systems with low energy consumption and real-time behavior since point clouds are the main data format directly obtained from most 3D sensors. Point clouds have higher structural complexity than 2D image data due to their disordered peculiarity, which means a lack of neighborhood consistency between data structures and spatial coordinates. The inconsistency leads to an irreproducible result when the convolution kernel is applied to raw point clouds without pre-processing. As a solution, [17, 36, 21] transforms and reorganize point clouds into voxels and extracts features using 3D convolution kernels. [6, 16] feed the neural networks with polygonal meshed spatial information as a substitute for the raw point clouds. However, these preprocessing approaches are not applicable in scenarios with real-time constraints and most of them are also not advantageous for semantic segmentation tasks. [23, 24] propose point cloud-applicable convolutional networks which concatenate the local features extracted by point-wise convolutional kernels with the global feature simply obtained by max-pooling layers and achieve the state-of-the-art accuracies on Modelnet40 [36], which is the currently most popular point cloud classification dataset and is also the one used in our experiments.

Explainability in 3D DNNs: Few studies have attempted to investigate the explainability of 3D DNNs. Although [39] refers to explainable point cloud classification, their work addresses the disorderly properties of point clouds using PointHop Units to adapt them to classical classifiers, which is part of the pre-processing rather than post-hoc explanations. [40] obtains point saliency maps by simply dropping points, which is not relevant to the explainability approaches. [12], the pioneer study of utilizing explainability approaches to point clouds remains crucial to our understanding of feature sparsity of 3D models. However, they only show sparse explanations that emphasize the importance of points at edges and corners, which is lack-of-semantics, and the evaluation criterion of the explanations is absent. In addition, the gradient-based methods are not adapted to models without gradients, such as tree-based models. In contrast, local surrogate model-based approaches are completely model-agnostic.

Explanation plausibility verification: Although there are many studies in the literature on the outcome of explainability methods, an acknowledged quantitative assessment for those approaches is absent [7] because explanations are subjective to humans. [1] argues that a feasible explanation should be sensitive to the weights of models and the data generating process, and proposed an alternative evaluation approach by randomizing the network weights as well as the labels and inspecting the sensitivity of the saliency maps. However, this approach tends to only benefit the gradient-based explainability methods and validates invalidity instead of feasibility. [11] strive to observe the improvement of the core performance of the network and the confidence they can generate for the users of the system when processing image data. [4, 27, 19] propose an intuitive and efficient pattern to verify the explanations by flipping the pixels that contribute positively or negatively (or approximately zero) to a particular class and record the verified prediction scores. Nevertheless, the flipping operation of this method could be optimized to some extent while processing point cloud data, which we will discuss in section 3.

3. EXPLAINABILITY APPROACHES FOR POINT CLOUDS

A significant advantage of surrogate model-based methods is that they are more widely applicable. In this section, we describe in detail our explainability approach, i.e. local surrogate model-based method for point cloud data based on LIME [26]. In addition, we elaborate the quantitative evaluation approach for point cloud explanations, which consider the local fidelity and plausibility of existing point cloud explainability methods. Note that as the surrogate model-based explainability approach is model-independent, the evaluation metrics are suitable for other data types besides point clouds as well.

3.1. Local surrogate model-based explainability approaches for Point Clouds

Local surrogate model-based explainability approaches aim to generate an explanation for a classifier f and a specific instance x from the data set X. To apply these methods to point cloud data, some pre-processing is necessary.

Algorithm 1 Pre-processing of Local surrogate model-based methods for point clouds

function 3D K-MEANS WITH FPS($P, n_c, maxIter$)
Input: $P \rightarrow N \times D$ point clouds, $n_c \rightarrow$ number of clusters, $maxIter \rightarrow$Max iterations
#Output indicates which cluster each point belongs to
Output: $C \rightarrow 1 \times N$ matrix
#Sample n_c points from P
$Centers \leftarrow$ FPS(c from P)
#Find the nearest cluster center for each point
while maxIter **do** :
for i in n_c **do**
$EDMatrix \leftarrow \|P, Centers\|_2$
$minDis \leftarrow \arg\min(EDMatrix)$
#Point belongs to the nearest cluster, update the centers
for j in n_c **do**
$P[C_j] \leftarrow$Where $minDis == j$
$newCenters \leftarrow$ Mean($P[C_j]$)
$Centers \leftarrow newCenters$
return C

3.1.1 Pre-processing with FPS

For explaining a point cloud input with size P utilizing local surrogate model-based explainability approaches, each point $p \in P$ is considered as a feature individually. However, to avoid explosive computational complexity and to organize the disordered point cloud data, we group the points into super-points C as features to be perturbed. We initialize a user-defined parameter n_c, indicating the number of clusters. To ensure uniformity and strengthen semantics we employ Farthest Point Sampling (FPS) to select n_c points from P and group all p according to spatial coordinates using 3D K-Means Clustering such that $\forall p \in P : p \in C_i$. The pseudo-code is presented in Algorithm 3.1.

3.1.2 Vanilla LIME applied for point clouds

Same as processing 2D images [26], LIME for point cloud data also satisfies the following constraint:

$$\xi(x) = \underset{g \in G}{argmin}\, L(f, g, \pi_x) + \Omega(g) \tag{1}$$

where f and g denote the classifier and the explainable model for a local instance x respectively, π_x denotes the proximity measure between samples z to the input x (locality around x), and $\Omega(g)$ denotes the complexity of the explainable model. LIME tries to minimize the locality-dependent loss $argminL(f, g, \pi_x)$ by approximating g to f. It takes samples z around x and feeds the perturbed samples z' into f to obtain a faithful surrogate model g that approximates f, also, it regularizes the complexity of the surrogate model g to guarantee that it is still explainable to humans.

As with most 2D image datasets, our 3D dataset for experiments has 40 different label categories, in which explainability is hardly guaranteed even for linear surrogate classifiers. Therefore, we train a linear regressor that approximates the prediction score of the corresponding category from the neural network. We sample $z \in Z$ by randomly flipping component clusters from x and feed the perturbed samples Z into the regressor g to obtain the predictions $g(z)$. To minimize $L(f, g, \pi_x)$, a kernel filters the generated samples Z around x based on the similarity between z and x proportionally (the fewer clusters being flipped the higher the weight). The surrogate model is subsequently trained with the weighted samples Z using linear regression. Due to the simplicity and transparency of linear models, it is explainable and understandable to humans and intuitive as to which parts (clusters) have positive/negative attributions to a particular prediction according to the parameters of the surrogate linear regressor.

3.1.3 Variable input size flipping (VISF)

LIME generates adjacent perturbation samples by flipping the corresponding clusters of the original instance. There are three widely-used flipping methods for a target cluster regarding 2D images: zero clearing all the included pixels, replacing those pixels with the average of the selected cluster (or the whole image), or reversing the sign of their coordinates. However, we argue that the above three operations can barely eliminate all information of the target cluster. For instance, although all pixel values are zeroed out, the contours formed by zeros remain on the data matrix and may still be learned by the neural networks. For a point cloud instance $P \in \mathbb{R}^{N \times D}$, the pixel values represent 3D spatial coordinates, and the above alternatives are likely to form a highly overlapping point set, resulting in uncertainty to determine whether the prediction fluctuations of the neural network are merely caused by flipping operations (see Fig S4 for intuitive visualizations).

To address the above problem, we simply discard the points contained in the target cluster c_i from the original instance as a means to completely ablate the information of the target cluster, i.e. $s_i = P \backslash c_i \in \mathbb{R}^{(N-\|c_i\|) \times D}$. This approach is only applicable for point cloud neural networks. Recall the architecture of point cloud networks, where the final symmetric function (i.e. the max-pooling layer) is used to extract the *global* feature from disordered point clouds while the *local* features in the lower layer are weighted by numerous 1×1 convolutional kernels, which allows the input size of the network to be arbitrarily reshaped without obstructing the inference. Notably, the variable input size flipping is both extendable in explaining and verification process (section 3.2.2).

3.1.4 Attribution summarizing

For explainability methods that return the importance of each spatial coordinate axis, the most popular and intuitive attribution summarization process is to simply sum them up:

$$C_p = \sum (C_1, C_2, C_3) \tag{2}$$

Where $C_{1\sim3}$ stand for the attributions in each of the three spatial axes. Different summarizing patterns have varied impacts on the explanations, which is worthy of further exploration.

3.2. Plausibility verification for 3D explanations

3.2.1 Local fidelity metrics

The fidelity indicates the prediction coherence between the original black-box model and the surrogate one, which is formulated as:

$$F = \frac{\sum \mathbb{1}\left(f(Z) = g(Z)\right)}{\|Z\|} \tag{3}$$

Nevertheless, instead of a classifier we utilize a linear regressor as the surrogate model, which returns the prediction score only associated with the predicted class. We thus compare the batched similarity between regression scores $g(Z) \in \mathbb{R}^{|Z|}$ and the prediction scores of the corresponding logits unit of the network $f(Z)$ via several loss and coefficient measurements:

- Mean loss: $L_m = \left| \sum_i^{\|Z\|} (\frac{f(Z)_i}{\|Z\|}) - \sum_i^{\|Z\|} (\frac{g(Z)_i}{\|Z\|}) \right|$
- Mean L1 and L2 loss: $L_1 = \sum_i^{\|Z\|} (\frac{|f(Z) - g(Z)|}{\|Z\|})$ and $L_2 = \sum_i^{\|Z\|} (\frac{(f(Z)_i - g(Z)_i)^2}{\|Z\|})$
- Weighted L1 and L2 loss: $L_1^\omega = \sum_i^{\|Z\|} (\frac{|f(Z)_i - g(Z)_i| \cdot \omega}{\|Z\|})$ and $L_2^\omega = \sum_i^{\|Z\|} (\frac{(f(Z)_i - g(Z)_i)^2 \cdot \omega}{\|Z\|})$
- Weighted coefficient of determination: $R_\omega^2 = 1 - \frac{\sum_i^{\|Z\|} (f(Z)_i - g(Z)_i)^2}{\sum_i^{\|Z\|} (f(Z)_i - \overline{f_\omega(Z)})^2}$

- Weighted adjusted coefficient of determination:

$$\hat{R}^2_\omega = 1 - (1 - R^2_\omega)\left[\frac{\|Z\|-1}{\|Z\|-\|g\|-1}\right]$$

where ω indicates the weights derived from the kernel, $\|Z\|$ denotes the number of observed samples, $\|g\|$ is the number of parameters of g and $\overline{f_\omega(Z)}$ indicates the weighted average. L_m,$L_1^{(\omega)}$ and $L_2^{(\omega)}$ measure the discrepancy in predicted scores while R^2_ω indicates the correlation between the prediction scores of the proxy model and the output of the neural network. In general, better agent approximations possess lower loss and higher decision coefficients with the predictions of neural networks. However, R^2_ω is sensitive to the number of samples and prone to positive bias under a small sample size [3]. We therefore introduce $\hat{R}^2_\omega$, which takes into account the size of variables and samples. Note that $\hat{R}^2_\omega$ has the meaningful range between $(-\infty, 1]$ under the assumption that $\|Z\| > \|g\|$, while the opposite case may exist in our experiments, it is therefore only referable for the case $\|Z\| > \|g\|$ in the experiment.

3.2.2 Method-independent explanation verification

Fidelity metrics are only suitable for surrogate model approaches. Additional measuring methodologies are required for the reliability of non-surrogate-based explainability methods such as gradient-based saliency maps. According to the hypothesis of the local accuracy of additive feature attribution [15]:

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i \tag{4}$$

the output of original model $f(x)$ is composed by linear summation of the individual feature attributions ϕ_i. One of the most intuitive ways to verify an explainability approach is to eliminate features with certain attributions ϕ_i (normally positive or negative) according to their generated explanations and observe whether the output of the model $f(x)$ exhibits corresponding variations:

$$f(x) - f(x\backslash i) \begin{cases} \geqslant 0 & \text{if } \phi_i x_i \geqslant 0 \\ \leqslant 0 & \text{if } \phi_i x_i \leqslant 0 \end{cases}, i \in M \tag{5}$$

where $\phi_i x_i$ denotes the attribution of flipped feature and $f(x\backslash i)$ denotes the output of the model after flipping feature i.

Nevertheless, in point cloud DNNs, the sensitivity of prediction scores for batch data is difficult to observe quantitatively due to the unevenly distributed prediction scores (the logits before the softmax) from different instances. Therefore, we normalize the variability of the predicted scores to facilitate its presentation in the form of an average prediction scoreline, which is formulated as

$$S_{avg} = \frac{1}{n}\sum_{i=0}^{n} \frac{S_i - S_{i_{min}}}{S_{i_{max}}} \tag{6}$$

where S_i denotes each score in the ith test (including positive, negative and random perturbation series), $S_{i_{min}}$ and $S_{i_{max}}$ denote the minimum and maximum values in the corresponding evaluation run respectively.

In addition, due to the use of clustered points, we determine the averaged attributions of clusters $\phi_c x_c$ in our work rather than of individual points $\phi_i c_i$, where

$$\phi_c x_c = \sum_{i=1}^{c} \phi_i x_i \tag{7}$$

, which lead to fluctuations in the prediction scores. The issue can be alleviated by increasing the number of clusters. We discuss this further in Section 4.

To quantitatively compare the plausibility among all types of explainability methods, we record the prediction scores of flipping the positive, negative and random contributing clusters respectively, denoted as S_{pos}, S_{neg} and S_{rdm}. The plausibility of the corresponding explanation can be formulated as:

$$\bar{p} = -\frac{\sum_i^{\|Z\|} \rho_i(S_{pos} - S_{rdm}, S_{neg} - S_{rdm})}{\|Z\|} \tag{8}$$

where $\rho(a, b)$ denotes the correlation coefficient between a and b. Intuitively, flipping positive clusters results in a decline of predicted scores while flipping negative clusters lifts them up. Flipping random clusters represents the impact of eliminating neutral clusters independent of attributions, as randomly selected clusters may consist of both positive and negative points and are therefore considered indifferent. $S_{pos} - S_{rdm}$ and $S_{neg} - S_{rdm}$ are then approximations of unbiased attribution-flipping processes. A plausible explanation should have exactly opposite sensitivities to contrary attributions, and therefore its correlation coefficient of the prediction score series is expected to be as small as possible, i.e., a high score of $\bar{p}$. We consider this value as a succinct description of the plausibility of the explainability method.

4. EXPERIMENT

In this section, we present the qualitative results of 3D surrogate model-based explainability methods (Sec. 4.1), evaluate and compare it with other 3D-applicable approaches utilizing the quantitative verifications proposed in section 3.2 (Sec. 4.2) and show how the explanations help

to analyse the samples classified incorrectly by the classifier. (Sec. 4.3). In our experiments[1], 1000 test instances are selected from Modelnet40 [36], which contains 12311 CAD models in 40 common categories and is currently the most widely-applied point cloud classification data set. We choose PointNet [23] as the model to be explained, which achieves an overall accuracy of 89.2% on Modelnet40. We sample 1024 points from each instance as input to the network. Additionally, we choose Exponential Smoothing kernel for training linear regressor, denoted as

$$K = \sqrt{e^{\frac{-d^2}{w^2}}} \tag{9}$$

where d denotes the distance from samples to the instances to be explained, and w denotes the kernel width which has an impact on the explanations. We therefore conduct a sensitivity experiment of the kernel widths from 0.05 to 0.3 and the corresponding results are demonstrated in section S1.1.2.

4.1. Qualitative explanation visualisation

Examples of explanations generated by PointNet, as well as their original point cloud structures are shown in Figure 1. What stands out in the figure is that explanations with different C are consistent overall except the one with $C = 1024$. We believe that the reason is that 1000 samples are insufficient for such a large number of clusters (1024) and therefore the surrogate model is not well-trained. Explanations based on clusters suffer from contribution neutralization. A cluster may consist of positive and negative contributing points simultaneously, aggregating them as an entity obscures the individual contribution of each point (C = 20,64 and 128). The neutralization can be alleviated by increasing the number of clusters, with the side effect of requiring more training samples and processing time ($C = 1024$).

4.2. Quantitative verification of explanation plausibility

Assessing the explanations by intuition is not quantitatively verifiable and is vulnerable to bias. This section mainly demonstrates the results of plausibility verification experiments i.e. local fidelity metrics in subsection 4.2.1 and the method-independent verification approach in subsection 4.2.2. There are two hyper-parameters for the proposed explainability method: Number of clusters C and number of perturbation samples S. In this section, we choose $C = 128$ and $S = 10^3$ as the standard performance of the proposed explainability method, since $C = 128$ is experimentally proven to achieve the best quantitative performance while maintaining the qualitative semantics. $S = 10^3$ generates high-qualified explanations within an acceptable processing time (see table S1) and thus is considered as the best configuration. Detailed experiments regarding hyper-parameters can be seen in Supplementary section S1.1.1.

[1]Our code is available at `https://github.com/Explain3D/LIME-3D`

4.2.1 Local fidelity metrics

Local fidelity metrics address measuring the prediction similarity between the original black-box model and the surrogate one, which play a pivotal role in verifying the plausibility of local surrogate model-based explainability methods. Due to the absence of related results as a reference, we treat the unmodified LIME (hard transplanted to point clouds) as the baseline. Table 1 compares the local fidelity of different explaining mechanisms, i.e. whether Farthest Points Sampling (FPS) is used or whether Variable Input Size Flipping (VISF) is employed. Corresponding metric symbols refer to section 3.2.1. According to the results, both FPS and VISF facilitate the improvement of local fidelity compared to the vanilla 3D LIME (baseline) as our LIME(FPS + VISF) improvement outperforms others in terms of most fidelity metrics. Note that the local fidelity only measures how closely the surrogate model approximates the black-box model. One drawback of the metric is that it is only applicable to explainability methods based on local surrogate models. Popular explainability methods (already proposed for point clouds) such as gradient-based ones are not compatible with these metrics, which confuses the user in choosing the most appropriate explainability method for specific tasks.

4.2.2 Method-independent plausibility verification

To address the aforementioned drawback we instead compare all existing point cloud-applicable explainability approaches utilizing the method-independent verification proposed in section 3.2.2. Again, we set $C = 128$ and $S = 10^3$ as the "competitor" of our proposed method. Besides positive and negative attributions, we also flip the same percentage of randomly-selected points as the baseline of prediction scores.

Figure 2 and Table 2 depict the trends of prediction scores and the correlation coefficient $\bar{p}$ between different existing 3D-applicable explainability methods. As the gradient-based approaches yield individual attributions for each point, we calculate coefficients for different percentages of points for fairness, i.e. top-$\%15$,$\%30$ and $\%50$ positive ones. What stands out in the results is that the explanations generated by 3D LIME and Integrated Gradients behave robustly. Their average prediction scores deteriorated rapidly after the gradual flipping of the most positive contribution points and conversely tended to increase when the negative contribution points are flipped.

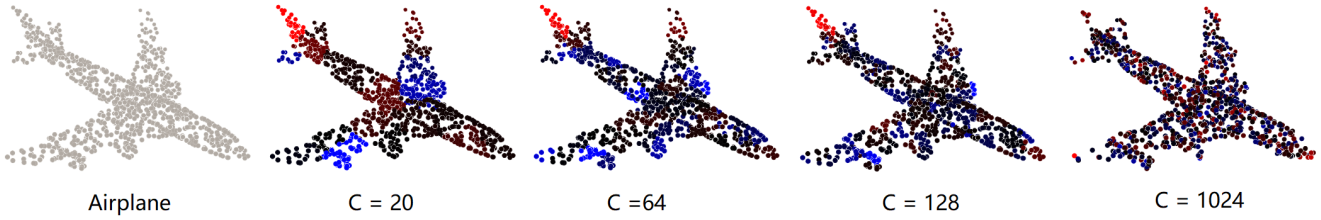

Figure 1. Examples of explanations with 1000 perturbation samples. C denotes the number of clusters. Brighter red points represent more positive contributions and, conversely, brighter blue points represent more negative contributions and dim points indicate zero contributions to the corresponding classification labels.

	L_m	L_1	L_1^ω	L_2	L_2^ω	R_ω^2	$\hat{R}_\omega^2$
LIME (baseline)	1.40×10^{-2}	1.11×10^{-1}	8.66×10^{-2}	1.06×10^{-1}	6.53×10^{-2}	0.338	0.241
LIME (FPS)	1.22×10^{-2}	9.80×10^{-2}	7.66×10^{-2}	8.67×10^{-2}	5.35×10^{-2}	**0.353**	**0.257**
LIME (VISF)	1.18×10^{-2}	1.01×10^{-1}	7.90×10^{-2}	9.68×10^{-2}	5.95×10^{-2}	0.335	0.237
LIME (FPS + VISF)	$\mathbf{1.03 \times 10^{-2}}$	$\mathbf{8.89 \times 10^{-2}}$	$\mathbf{6.95 \times 10^{-2}}$	$\mathbf{7.84 \times 10^{-2}}$	$\mathbf{4.82 \times 10^{-2}}$	0.346	0.249

Table 1. Local fidelities of different explaining mechanics for point cloud data, where FPS denotes employing Farthest Point Sampling instead of randomly choose clusters and VISF denotes the Variable input size flipping mechanism. The unmodified application of LIME to point clouds is regarded as the baseline.

	$\bar{p}_{.15}$	$\bar{p}_{.3}$	$\bar{p}_{.5}$
Vanilla Gradients	-0.574	-0.569	-0.672
Guided Back-propagation	-0.741	-0.695	-0.623
Integrated Gradients	0.484	0.366	0.236
KernelSHAP	-0.205	-0.257	-0.256
LIME (FPS + VISF)	**0.622**	**0.531**	**0.372**

Table 2. Plausibility $\bar{p}$ of flipping top-%15,%30 and %50 attributed points.

	$\bar{p}_{.15}$	$\bar{p}_{.3}$	$\bar{p}_{.5}$
LIME (baseline)	0.598	0.520	0.341
LIME (FPS)	0.615	0.513	0.361
LIME (VISF)	0.593	0.514	0.345
LIME (FPS + VISF)	**0.622**	**0.531**	**0.372**

Table 3. Plausibility of different explaining mechanics for enhancing the explanation quality.

On the other hand, Vanilla Gradients, Guided Back-propagation and 3D KernelSHAP are unable to distinguish between points with different contributions, resulting in gradient maps being less uniform than Integrated Gradients [12]. Interestingly, KernelSHAP is a variant of LIME based on Shapley value, differing from the latter solely in the choice of kernels. KernelSHAP assigns high weights to perturbation samples with only a minority of clusters remained, which severely impairs the global structure of the instance. Empirically, we find that such kernels may be more suitable for black-box model structures on other data types, but with limited performance in explaining point clouds.

We also compare the plausibility among all explanation mechanisms, the corresponding scores are presented in table 3. The proposed method also dominates which is consistent with the results in section 4.2.1.

4.3. Applying local surrogate model-based explainability methods for failure analysis

A potentially applicable prospect of local surrogate model-based explainability methods is failure analysis. This analysis has important implications for understanding the erroneous attention paid by the classifier and provides opportunity for further research, e.g. 3D model revision. Figure 3 shows examples of the attributions to the misclassified instances.

As can be seen from Figure 3, a majority of the misclassifications were caused by misdirected attention. In the first two examples, the faucet instead of the bathtub itself possesses the most positive attribution, misleading the model to neglect the structure of the primary target object. We believe this is because only a tiny fraction of the bathtubs in the training data is accompanied by a faucet. Another similar type of error frequently occurs with the class "Flower pot". The major attention of the model is drawn to the plant above rather than the pot below, resulting in a prediction of 'Plant' instead of "Flower pot" (the pot even draws a negative contribution to the ground truth label). From a human perspective, this type of data is ambiguously labeled, as both classes "Plant" and "Flower pot" are reasonable ground truth. Towards a more accurate model, this labeling type should be avoided whenever possible.

5. CONCLUSION

Although point cloud neural networks have received critical attention in recent years, so far, there have been few studies on their explainability. Our work proposes an ex-

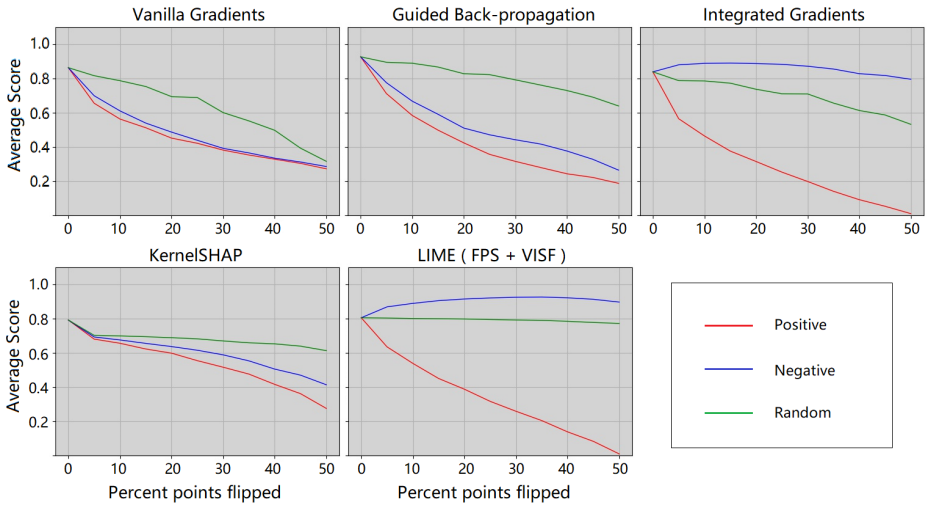

Figure 2. Variation trends of the prediction scores (y-axis) by flipping and re-inference. The scores are the average of normalized prediction scores of 1000 test instances. Red and blue lines indicate the trend of flipping positive and negative contribution points, respectively, the green line indicates flipping random points that are independent of contribution. The x-axis indicates the percentage of flipped points for a given instance.

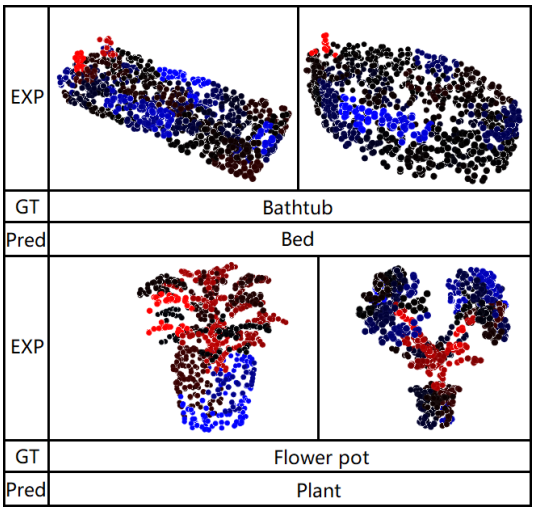

Figure 3. Explanation of the misclassified examples. Brighter red points indicate more positive contributions, while brighter blue points indicate more negative contributions and dim points indicate zero contributions. All contributions are concerning the prediction class (wrong class instead of the ground truth).

plainability approach for point clouds based on LIME [26]. We also provided the possibility to quantitatively validate the point cloud explanations. We evaluated and compared the performance of our approach against different existing explainability methods for point cloud data. The evaluation comparison revealed that our local surrogate model-based approaches as well as Integrated Gradients yield relatively plausible explanations and outperform other methods such as Guided Back-propagation. Our results also demonstrated that a larger amount of clusters and more perturbed samples are required to avoid compromising fidelity, which however consumes more processing time. Moreover, we provided intuitive analyses for misclassified samples by utilizing the proposed method. The analyses showed that part of the misclassified cases can be attributed to the anomalous structural distributions or ambiguous labels of the input data, misdirecting the attention of the classifier.

This work attempted to shed light on 3D neural networks. There is still tremendous potentials for further progress. Most local surrogate model based-explainability methods suffer from sample distortion as they treat each feature independently, with neither spatial relationships nor causality taken into consideration, producing unlikely feature combinations and resulting in reduced quality of the explanation. Constraining the causal structure of perturbed samples to more closely resembling the training samples by introducing prior knowledge is a promising idea. Another potential area of research is to generate more comprehensible and interesting explanations for point clouds, for instance, global explainability approaches or instance-based methods such as activation maximization or adversarial examples.

6. Acknowledgment

This research has been funded by the Federal Ministry of Education and Research of Germany as part of the competence center for machine learning ML2R (01IS18038A).

References

[1] Julius Adebayo et al. Sanity checks for saliency maps, 2020. arXiv preprint, arXiv:1810.03292.

[2] Md Manjurul Ahsan et al. Study of Different Deep Learning Approach with Explainable AI for Screening Patients with COVID-19 Symptoms: Using CT Scan and Chest X-ray Image Dataset. 2020. arXiv preprint, arXiv:2007.12525.

[3] M.A. Ali. Effect of sample size on the size of the coefficient of determination in simple linear regression. *Journal of Information and Optimization Sciences*, 8(2):209–219, 1987.

[4] Sebastian Bach et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):1–46, 2015.

[5] Alejandro Barredo Arrieta et al. Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58(October 2019):82–115, 2020.

[6] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs. pages 1–14, 2013. arXiv preprint, arXiv:1312.6203.

[7] Nadia Burkart and Marco F. Huber. A Survey on the Explainability of Supervised Machine Learning. pages 1–74, 2020. arXiv preprint, arXiv:2011.07876.

[8] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data, 2018. arXiv preprint, arXiv:1808.02610.

[9] Luca Cultrera, Lorenzo Seidenari, Federico Becattini, Pietro Pala, and Alberto Del Bimbo. Explaining autonomous driving by learning end-to-end visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[10] Andre Esteva et al. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

[11] Van Gemert. Evaluating the performance of the LIME and Grad-CAM explanation methods on a LEGO multi-label image classification task. 2020. arXiv preprint, arXiv:2008.01584v1.

[12] A. Gupta, S. Watson, and H. Yin. 3d point cloud feature explanations using gradient-based methods. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[13] Markus Hofmarcher et al. *Visual Scene Understanding for Autonomous Driving Using Semantic Segmentation*, pages 285–296. Springer International Publishing, Cham, 2019.

[14] B. Kim et al. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4149–4157, 2019.

[15] Scott M. Lundberg and Su In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-Dec.(Section 2):4766–4775, 2017. arXiv preprint, arXiv:1705.07874.

[16] Jonathan Masci, Davide Boscaini, Michael M. Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on Riemannian manifolds. pages 37–45, 2015. arXiv preprint, arXiv:1501.06297.

[17] Daniel Maturana and Sebastian Scherer. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. *Iros*, pages 922–928, 2015.

[18] Pablo Messina et al. A survey on deep learning and explainability for automatic image-based medical report generation, 2020. arXiv preprint, arXiv:2010.10563.

[19] Grégoire Montavon. *Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison*, pages 253–265. Springer International Publishing, Cham, 2019.

[20] Anh Nguyen, Jason Yosinski, and Jeff Clune. *Understanding Neural Networks via Feature Visualization: A Survey*, pages 55–76. Springer International Publishing, Cham, 2019.

[21] Charles R. Qi et al. Volumetric and Multi-View CNNs for Object Classification on 3D Data. 2016. arXiv preprint, arXiv:1604.03265.

[22] Charles Ruizhongtai Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum PointNets for 3D Object Detection from {RGB-D} Data. *CoRR*, abs/1711.0, 2017. arXiv preprint, arXiv:1711.08488.

[23] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[24] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. 2017. arXiv preprint, arXiv:1706.02413.

[25] S. Ramos, S. Gehrig, P. Pinggera, U. Franke, and C. Rother. Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1025–1032, 2017.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:1135–1144, 2016. arXiv preprint, arXiv:1602.04938v3.

[27] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

[28] Wojciech Samek and Klaus-Robert Müller. *Towards Explainable Artificial Intelligence*, pages 5–22. Springer International Publishing, Cham, 2019.

[29] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences, 2019. arXiv preprint, arXiv:1704.02685.

[30] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. arXiv preprint, arXiv:1312.6034.

[31] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. arXiv preprint, arXiv:1706.03825.

[32] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015. arXiv preprint, arXiv:1412.6806.

[33] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Gradients of counterfactuals, 2016. arXiv preprint, arXiv:1611.02639.

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, 7:5109–5118, 2017. arXiv preprint, arXiv:1703.01365.

[35] Tom Vermeire and David Martens. Explainable image classification with evidence counterfactual, 2020. arXiv preprint, arXiv:2004.07511.

[36] Zhirong Wu et al. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[37] Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, and Sally Shrapnel. Deep neural network or dermatologist? In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 48–55, Cham, 2019. Springer International Publishing.

[38] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*, pages 2018–2025, 2011.

[39] Min Zhang, Haoxuan You, Pranav Kadam, Shan Liu, and C.-C. Jay Kuo. Pointhop: An explainable machine learning method for point cloud classification. *IEEE Transactions on Multimedia*, 22(7):1744–1755, Jul 2020.

[40] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Pointcloud saliency maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[41] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.