EG UK Computer Graphics & Visual Computing (2024)
A. Slingsby and D. Hunter (Editors)

Assessing the Reliability of Integrated Gradients-Based Saliency Maps for 3D Point Cloud Semantic Segmentation Models

J. F. Ciprián-Sánchez[1], J-M. Burmeister[2], T. Cech[2], R. Richter[2] and J. Döllner[1]

[1]University of Potsdam, Digital Engineering Faculty, Hasso Plattner Institute, Germany
[2]University of Potsdam, Digital Engineering Faculty, Germany

Abstract

Deep learning models achieve high accuracy in the semantic segmentation of 3D point clouds; however, it is challenging to discern which patterns a model has learned and how it derives its output from the input. Recently, the Integrated Gradients method has been adopted to explain semantic segmentation models for 3D point clouds. This method can be used to generate saliency maps that visualize the contribution of input points to a particular model output. However, there is a lack of quantitative evaluation of the reliability of the generated saliency maps and the influence of the baseline selection (a central component of Integrated Gradients) on the method's results. In this paper, we quantitatively evaluate the reliability of saliency maps generated by the Integrated Gradients method for a 3D point cloud semantic segmentation model through well-known sanity checks from the image domain that we adapt to 3D point cloud segmentation. We perform these sanity checks for three different baselines to further evaluate the stability of the generated saliency maps concerning the baseline choice. Our results indicate that the Integrated Gradients method is sensitive to both the parameters of the model and training labels, unstable concerning the choice of baseline, and that, although it can identify points with high contributions to the model output, it fails to identify correctly if such contributions are positive or negative. Finally, we propose an averaging approach to aggregate the results of points that receive multiple scores from Integrated Gradients during the segmentation process and show that it produces saliency maps that better reflect high-contribution input points than previous approaches.

CCS Concepts

• ***Human-centered computing*** → ***Visualization design and evaluation methods;*** *Geographic visualization;* • ***Computing methodologies*** → ***Neural networks;***

1. Introduction

3D point clouds are widely used in geospatial applications, as they can serve as point-based 3D models or base data for 3D model reconstruction. Semantic segmentation, which aims to assign an object class label to each point, plays a fundamental role in a growing number of geospatial applications [WK19; GHD*17]. Deep learning (DL) models achieve high accuracy in semantic segmentation of 3D point clouds; however, the large number of parameters and layers of DL models makes it challenging to discern which patterns a model has learned and how it derives its output from the input data. Attributing the predictions of a DL model to its input would be beneficial to satisfy regulatory requirements, debug models, or verify whether a model has learned unintended patterns [AGM*18]. To address this issue, several methods for generating saliency maps have been proposed for image classification [LBB*23]. These methods aim to generate saliency maps that visualize how different pixels of an image contribute to the prediction score for a given class. Although some of these methods have been transferred or adapted to 3D point cloud classification [TSS23; ZCY*19; ZWQF20; GWY20], transferring these approaches to 3D point cloud semantic segmentation is not straightforward. Since semantic segmentation can be considered a point-wise classification problem, a naïve approach would be to generate separate saliency maps for each point. However, this would result in a high number of saliency maps, which would not be feasible for visual inspection. Therefore, approaches for aggregating the saliency maps of individual points are required. Furthermore, DL models for semantic segmentation of large-scale 3D point clouds usually rely on dividing the 3D point clouds into smaller subsections and processing them separately. This raises the additional question of how saliency maps for individual subsections can be aggregated, especially if the subsections overlap.

The work by SCHWEGLER, MÜLLER, and REITERER [SMR23] is one of the first to explore these issues, applying the Integrated Gradients (IG) method to explain 3D point cloud semantic segmentation models. In the IG method, a range of inputs is generated by linearly interpolating along a straight-line path between a baseline input and the original input, and the gradients of the input

DOI: 10.2312/cgvc.20241217

features integrated over all generated inputs are used as a measure of saliency [STY17]. In the approach of SCHWEGLER, MÜLLER, and REITERER [SMR23], aggregated saliency maps are computed by combining the prediction scores of all points in a class and keeping only the last attribution when a point is contained in multiple subsections. However, their work only presents the results visually and lacks a quantitative evaluation of whether the obtained results correctly reflect the relevance of the input for the model. Furthermore, they do not evaluate how the saliency maps are influenced by the choice of baseline, although this has been shown to have a strong influence on the resulting saliency maps for image classification models [HSM*22].

Therefore, our work focuses on a quantitative evaluation of IG-based saliency maps for the semantic segmentation of large-scale 3D point clouds, using several previously proposed metrics to evaluate saliency maps [AGM*18; PDS18; TSS23]. Specifically, three sanity checks are conducted on IG-based saliency maps obtained from the RandLA-Net architecture [HYX*20] trained on the PARIS-CARLA-3D dataset [DDR*21]: (1) Using the model parameter randomization and (2) the data randomization tests proposed by ADEBAYO, GILMER, MUELLY, et al. [AGM*18] it is tested whether the saliency maps obtained from IG are sensitive to the parameters of the model to be explained and the training labels. (3) Using ablation tests based on the work of PETSIUK, DAS, and SAENKO [PDS18], it is tested whether gradually removing the input points with the highest and lowest attribution scores (i.e., saliency) decreases the average prediction probability of the remaining points for a given class. To test whether the IG-based saliency maps are stable with respect to the baseline choice, the sanity checks are conducted for three different baselines. Furthermore, we propose to average the obtained attributions of points present in multiple subsections of the point cloud, instead of the overwriting approach proposed by SCHWEGLER, MÜLLER, and REITERER [SMR23], assessing the reliability of both approaches through the aforementioned sanity checks across all evaluated baselines.

Our results show that the IG method for semantic segmentation of 3D point clouds is sensitive to the model parameters and the relationship between the observed data and the training labels. Furthermore, it shows instability with respect to the studied baselines, showing different results in the ablation tests and saliency maps for each of them. Through the ablation tests, we find that the absolute values of the IG output better reflect the model behavior, suggesting that IG fails to correctly identify whether the contributions of points towards the model output are positive or negative. Finally, we find that our proposed averaging approach performs better in the ablation tests and produces visually clearer saliency maps.

2. Related Work

2.1. Saliency Explanation Methods for 3D Point Cloud Models

The use of explainable artificial intelligence (XAI) techniques to explain DL models for 3D point cloud analysis is a relatively new field, with initial works exploring the use of saliency methods [GWY20; MPF*22; TSS23; ZCY*19; ZWQF20], surrogate models [TK22], activation maximization [Tan23b], and point attacks [TK23]. Our work focuses on saliency methods, i.e., methods that aim to visualize the contribution of different input features to a particular model output. Most existing work on saliency methods for 3D point cloud models targets classification tasks. For example, GUPTA, WATSON, and YIN [GWY20] transfer several gradient-based saliency methods from the image domain to 3D point cloud classification, namely vanilla gradients [SVZ14], Guided Backpropagation [SDBR15], and IG [STY17]. The authors evaluate them on two DL architectures and find that the methods studied assign high attribution values to edges and corners, with IG coupled with PoinetNet [QSMG17] producing more uniform saliency maps. [MPF*22] implement the Grad-CAM [SCD*17] algorithm for 3D point cloud classification models and combine it with visualizations of a model's latent features based on dimensionality reduction. ZIWEN, WU, QI, and FUXIN [ZWQF20] generate saliency maps by optimizing a loss function that aims to maximize the predicted probability for a target class when the saliency map is used as an input mask and to minimize it when the inverse saliency map is used as an input mask. The masking procedure is implemented using a curvature-based smoothing approach. ZHENG, CHEN, YUAN, et al. [ZCY*19] suggest constructing saliency maps for 3D point cloud classification models by point dropping. Since point dropping is a non-differentiable operator, they approximate point dropping by shifting points towards the point cloud centroid and use the gradients of the classification loss with respect to the point radius in a spherical coordinate system as a measure of saliency. TAYYUB, SARMAD, and SCHÖNBORN [TSS23] propose a saliency mapping technique that combines gradient information with point dropping named Accumulated Piece-Wise Explanations (APE).

To the best of our knowledge, the works by KURIYAL and KUMAR [KK24] and SCHWEGLER, MÜLLER, and REITERER [SMR23] are the only ones that study saliency methods for 3D point cloud semantic segmentation models. KURIYAL and KUMAR [KK24] propose point Grad-Seg Class Activation Mapping (pGS-CAM), a saliency method that extends GradCAM to semantic segmentation by aggregating the per-point gradients via summation. The authors generate saliency maps that represent the feature importance for the segmentation results aggregated over an entire point cloud and over subsets of points.

SCHWEGLER, MÜLLER, and REITERER [SMR23] apply the IG [STY17] approach to point cloud semantic segmentation. They visualize the explanations generated by IG for three target classes to evaluate the impact of coordinate and color features on the semantic segmentation performance. They use a data processing pipeline for large-scale 3D point clouds that divides the point clouds into multiple overlapping subsets and processes each subset by a DL model. Given this pipeline, the authors compute aggregated saliency maps by combining the prediction scores of all points in a class, and keeping only the last attribution of a point when it is contained in multiple subsets. Since SCHWEGLER, MÜLLER, and REITERER [SMR23] do not provide a quantitative evaluation on the reliability of the generated saliency maps, nor their stability with respect to the choice of baseline, our work focuses on the quantitative evaluation of the IG method for point cloud semantic segmentation.

2.1.1. Integrated Gradients

IG is a saliency method that generates a set of inputs by linear interpolation along a straight-line path between a baseline input and the

original input to be explained. Then, the gradients of the model output with respect to its input are integrated over all generated inputs and used as a measure of saliency. IG satisfies the properties of sensitivity and implementation invariance. The sensitivity property is satisfied if, for every input and baseline that differ in one feature but have different predictions, the differing feature receives a non-zero attribution. Implementation invariance signifies that if two models are functionally equivalent, their corresponding attributions must be identical [STY17; HSM*22].

The choice of baseline depends upon the domain and task and has been shown to impact the obtained attributions [KHA*19]. Different types of baselines have been explored, especially in computer vision and text analysis; some of the most common baseline approaches include black images or zero vectors [STY17], white images [PDS18], randomly initialized baselines, maximum-distance to input and blurred baselines [SLL20] and, more recently, baselines that maximize the entropy of the classification logits [Tan23a]. Although there is initial work on the use of IG or IG-based approaches for the explanation of 3D point cloud semantic segmentation [SMR23] DL models, to the best of our knowledge, the impact of the baseline selection on the obtained attributions has only been studied for the image domain [SLL20].

2.2. Sanity Checks and Metrics for Saliency Map Evaluation

One of the main challenges of saliency methods is that, given the lack of ground truth, it is difficult to assess the quality of the generated saliency maps [AGM*18]. While some authors evaluate the usefulness of saliency maps through user studies [HHH22], several sanity checks and metrics have been proposed for quantitative evaluation: ADEBAYO, GILMER, MUELLY, et al. [AGM*18] propose two sanity checks for saliency maps: a model parameter randomization test and a data randomization test. The model parameter randomization test compares saliency maps from a trained and a randomly initialized model. The data randomization test compares saliency maps from a model trained with correctly labeled data and a model trained with randomly permutated labels. Since saliency methods should depend on the learned parameters of a model and the labels of the training instances, the saliency maps should be substantially different in both cases.

HEDSTRÖM, WEBER, LAPUSCHKIN, and HÖHNE [HWLH24] re-examine the model parameter randomization test and propose two variations of it: Smooth Model Parameter Randomisation Test (sMPRT) and Efficient Model Parameter Randomisation Test (eMPRT). sMPRT aims to mitigate the impact of noise through a de-noising step. eMPRT replaces pairwise similarity measures (Structural Similarity Index Measure (SSIM) and Histogram Of Oriented Gradients (HOG)) with an entropy-based complexity metric to better evaluate the similarity between the test results and the evaluated saliency maps. KINDERMANS, HOOKER, ADEBAYO, et al. [KHA*19] suggest testing whether adding a constant shift to the input data, which is then reversed within the model, changes the results of a saliency method. They show that saliency methods for image classification can produce misleading attributions since they are sensitive to transformations that do not affect the model performance.

In another line of work, sanity checks have been proposed that remove or modify the input features with the highest or lowest attribution values and measure how this affects the probability predicted for a target class. For example, STURMFELS, LUNDBERG, and LEE [SLL20] employ a top-K ablation test and a mass center ablation test. The top-K ablation test evaluates whether removing the top K features with the highest attribution values reduces the predicted output logits for the target class [SLL20]. The mass center ablation test [GAZ19] calculates the center of mass of the saliency map and removes a region around this center to determine whether the saliency map highlights an important region in the image [SLL20]. Similarly, PETSIUK, DAS, and SAENKO [PDS18] propose a removal metric that measures the decrease in the probability predicted for the target class while gradually deleting or masking the pixels with the highest attribution. Additionally, they propose an insertion metric that measures the increase in the probability predicted for the target class as the pixels with the highest attribution are gradually added to a baseline image. In the context of 3D point cloud classification, TAYYUB, SARMAD, and SCHÖNBORN [TSS23] use a point-dropping approach to progressively remove points from the 3D point clouds based on their attribution values, measuring the decrease in classification accuracy during the process. The authors use two variations: high-drop, i.e., removing high-relevance points first, and low-drop, i.e., discarding low-relevance points first. KURIYAL and KUMAR [KK24] as well as ZHENG, CHEN, YUAN, et al. [ZCY*19] also use point-dropping approaches to evaluate the quality of saliency maps for different point cloud classification architectures.

3. Data and Methods

We conduct our experiments using the RandLA-Net architecture [HYX*20], a multi-layer perceptron (MLP)-based architecture specifically designed to segment large-scale 3D point clouds. Due to its fast inference times, it allows for the incorporation of XAI techniques while keeping the total execution times within feasible boundaries. For the IG implementation, we use PyTorch's *autograd* engine; it computes the vector-Jacobian product to obtain the gradients of the predicted class probability with respect to the model input averaged over all points of the model input. Based on the findings of SCHWEGLER, MÜLLER, and REITERER [SMR23], we use seven interpolation steps for the IG computation. Appendix A details the hardware and software used for all experiments.

3.1. Paris-CARLA-3D Dataset

The PARIS-CARLA-3D dataset [DDR*21] consists of two subsets of 3D point clouds of outdoor environments. One of the subsets was synthetically generated and contains 700 million points. The second subset contains 3D point clouds acquired in Paris, France, and consists of 60 million points. We only use the subset acquired in the city of Paris, as it provides real-world data while allowing for feasible experiment run times. We refer to this subset of the PARIS-CARLA-3D dataset as the PARIS dataset in subsequent sections of this paper. For training, validation, and testing, we follow the partitions proposed in the original paper [DDR*21], using the *Soufflot0* partition for the evaluation of the generated saliency maps. We perform all experiments with both the signed attribution scores and

their corresponding absolute values for three target classes: *Building*, *Vegetation*, and *Roadline* (i.e., road markings), as they represent objects with very different geometries, colors, and sizes.

3.2. Data Preprocessing

Following a similar approach as THOMAS, QI, DESCHAUD, et al. [TQD*19], we initially reduce the point density of the PARIS dataset through grid sampling. We set the grid size to 6 cm and use point coordinates and RGB values as model input. In line with the work of KUMAR, ANDERS, WINIWARTER, and HÖFLE [KAWH19], we calculate normal vectors and curvature values for each point and add them to the input features. The normal vector of a point is estimated by calculating the eigenvectors of the 3D covariance matrix of a point's k_n nearest neighbors. The eigenvector with the smallest eigenvalue is taken as the normal vector [HDD*92]. To calculate the curvature value of a point, a tangent plane is spanned by the point and its normal vector. The curvature value is defined as the average distance of the point's k_c nearest neighbors to this tangent plane [PGK02]. In this work, we use empirically selected values of $k_n = 78$ and $k_c = 16$. We manually remove high-curvature outliers using the CloudCompare software.

Due to memory constraints, we use a data processing pipeline in which spherical neighborhoods of fixed spatial extent (6 m radius) are randomly sampled from the large-scale 3D point clouds and processed separately by the DL models. Furthermore, each neighborhood is thinned to a fixed number of points (4096) through random sampling, as this simplifies the batch processing of the data and ensures that the DL model can process a batch of samples with a given GPU memory budget. During inference, we sample model inputs from a large-scale 3D point cloud until every point has received at least one prediction. If points are contained in multiple neighborhoods, the predictions are averaged.

3.3. Handling of Points with Multiple Attributions

Our data preprocessing pipeline described in Section 3.2 can result in points receiving more than one prediction and attribution score. To construct the final saliency map of the whole 3D point cloud, SCHWEGLER, MÜLLER, and REITERER [SMR23] address this problem by overwriting the point-wise attribution scores with the most recent ones on each iteration. Given that the final prediction score for a given point is computed by averaging all the prediction scores it received throughout the process, for consistency, we propose to average the obtained attributions as well. Intuitively, points that obtain high attribution values consistently over multiple neighborhoods should contribute more to the prediction score for a given class. We visualize the saliency maps obtained through our approach and the one by SCHWEGLER, MÜLLER, and REITERER [SMR23] and evaluate their reliability through the sanity checks and baseline tests described in Section 3.5 to Section 3.7.

3.4. Saliency Map Visualization

To visualize all obtained saliency maps, congruent with existing work [STY17; SMR23], we first obtain a single attribution score for each point by aggregating the attributions of all per-point features. We use two variations of this aggregation: keeping the original attribution signs and using their corresponding absolute values.

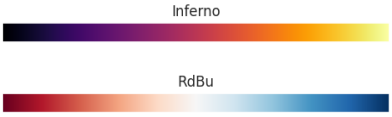


Figure 1: *Sample of the colormaps used, illustrating their variation for values in the* $[0, 1]$ *range.*

We use a perceptually uniform sequential color map for the absolute value attributions, i.e., one in which the lightness value increases monotonically through the color map - appropriate for representing ordered data. To visualize the signed attribution values, we use a divergent color map (appropriate for representing data that deviates around 0), i.e., one in which the lightness value increases monotonically up to a maximum, followed by monotonically decreasing values. We use Matplotlib's *inferno* sequential and *RdBu* divergent color maps [mat24], which are shown in Fig. 1.

As SMILKOV, THORAT, KIM, et al. [STK*17] mention, visualizing gradient-based saliency maps is surprisingly nuanced - the presence of outliers with much higher attribution values than the majority of the points (in our case, due to exploding gradients during the IG computation) can strongly affect the color mapping and have a large impact on the resulting visualization. To address this issue, we clip the attribution values using minimum and maximum thresholds. To identify appropriate clipping thresholds, we plot the distribution of the absolute value attributions (see Appendix B for further details). After clipping, we normalize all attribution values to the $[0, 1]$ range for their mapping to the corresponding color map. For the saliency maps obtained in the randomization tests, we use the same clipping thresholds as their corresponding original saliency maps and use them for the normalization and color mapping to allow for an adequate visual comparison.

3.5. Randomization Tests

To evaluate the sensitivity of the IG method to the model parameters and the training data, we conduct the model parameter and the data randomization tests proposed by ADEBAYO, GILMER, MUELLY, et al. [AGM*18] on the saliency maps obtained by both our attribution-averaging approach and the overwriting approach by SCHWEGLER, MÜLLER, and REITERER [SMR23]. ADEBAYO, GILMER, MUELLY, et al. [AGM*18] use the SSIM and HOG image similarity metrics together with the Spearman rank correlation score to assess the similarity of the saliency maps obtained from the randomization tests and the original model. Given that it is not straightforward to transfer image comparison methods, we rely on the Spearman rank correlation metric to assess the test results. To avoid introducing distortion in the test results due to the conversion to an RGB saliency map (see Section 4.3), we compute the Spearman rank correlation score directly on the attribution values.

As a point of comparison for the obtained results and in line with [AGM*18], we compute the Spearman rank correlation score

between (1) the original saliency maps and randomly generated ones, and (2) between pairs of randomly generated saliency maps, in both cases using a random uniform distribution.

3.6. Ablation Tests

Similar to [PDS18] and [TSS23], we perform ablation tests to evaluate how the removal of high- and low-attribution points impacts the average prediction probability for three target classes. We perform the ablation tests by progressively removing up to 80% (with a step size of 10%) of high- and low-attribution points at two levels of granularity: per neighborhood and on the final saliency maps for the entire 3D point cloud. For the per-neighborhood evaluation, we randomly sample 500 neighborhoods from the original 3D point cloud, compute the attribution scores, and remove the selected percentage of points from the neighborhood. Since our pipeline requires a fixed input size of 4096 points, we randomly duplicate the remaining points until the required input size is reached. We then compute the average prediction scores for the target class across all points in the neighborhood and average them across all sampled neighborhoods, excluding the duplicated points. To evaluate the saliency maps for the entire 3D point cloud, we directly remove the high and low attribution points from the *Soufflot0* partition, using the ablated dataset as input. We then present the average predicted probability values for each class over the entire 3D point cloud.

3.7. Comparison of Baseline Inputs

We compare the random and zero-vector baselines, which are commonly used for IG [SLL20; STY17], and the max-entropy baseline, which was recently proposed by TAN [Tan23a]. For the case of 3D point clouds, we define the random baseline as a random uniform distribution of points in a volume with the same shape and spatial extent as the original model inputs (see Section 3.2). We produce a random distribution of points over a spherical geometry with a 6 m radius for the x, y, and z coordinates while generating random values for all additional per-point features within their minimum and maximum possible values. The max-entropy baseline is defined as follows [Tan23a]:

$$B_{Xentr} = \arg\max_{x} H(\text{Softmax}(f_l(x))), \tag{1}$$

where f_l are the logits of the model and $H(*)$ is the Shannon entropy function [Tan23a]. We start with a random baseline and optimize it following Eq. (1) via gradient ascent over 100 epochs, keeping the baseline with the highest entropy. Finally, for the zero-vector baseline, we initialize all per-point features to zero.

4. Results

4.1. Randomization Tests

Concerning the reference values for the randomization tests (see Section 3.5), for the first case (output saliency map compared to a randomly generated one), we obtain an average Spearman rank correlation score of -1.281×10^{-5} with a standard deviation $\sigma = 0.00039$, with its corresponding p-values showing an average of 0.5481, with $\sigma = 0.2851$. For the second one (comparing randomly generated saliency maps), we get a Spearman rank correlation score of 1.7948×10^{-5} with $\sigma = 0.0003$ and p-values of 0.4796 with $\sigma = 0.2543$. These correlation scores indicate that the saliency maps are non-significantly correlated, while the higher p-values indicate that the null hypothesis (i.e., that the correlation is due to chance) cannot be rejected - which is because we are comparing with randomly generated data. In contrast, successful randomization test results would have a low Spearman rank correlation score and a low p-value, indicating that the saliency maps are weakly correlated (i.e., significantly different) and that this is unlikely to be due to chance. Table 1 and Table 2 show low correlation scores between the original saliency maps and the ones resulting from model and data randomization, indicating no evidence that IG is insensible to model parameters and training data. Figs. 12 to 14 (in Appendix C) show examples of saliency maps obtained in the randomization tests for all classes.

Class	Baseline	Signed attribution		Abs. attribution	
		Score	p-val.	Score	p-val.
Building	Random	0.013	0.0	−0.019	0.0
	Max-entropy	0.128	0.0	−0.011	0.0
	Zero-vector	0.113	0.0	0.189	0.0
Roadline	Random	0.033	0.0	0.130	0.0
	Max-entropy	0.021	0.0	0.1	0.0
	Zero-vector	0.037	0.0	−0.238	0.0
Vegetation	Random	0.089	0.0	0.041	0.0
	Max-entropy	0.078	0.0	−0.002	0.0
	Zero-vector	−0.037	0.0	0.042	0.0

Table 1: *Spearman rank correlation scores between the original attributions and those from the model parameter randomization test.*

Class	Baseline	Signed attribution		Abs. attribution	
		Score	p-val.	Score	p-val.
Building	Random	−0.009	0.0	−0.097	0.0
	Max-entropy	0.01	0.0	−0.044	0.0
	Zero-vector	0.02	0.0	−0.049	0.0
Roadline	Random	0.002	0.0	0.1	0.0
	Max-entropy	0.001	0.0	0.059	0.0
	Zero-vector	0.022	0.0	0.106	0.0
Vegetation	Random	0.002	0.0	0.009	0.0
	Max-entropy	0.001	0.0003	0.002	0.0
	Zero-vector	0.004	0.0	0.038	0.0

Table 2: *Spearman rank correlation scores between the original attributions and those from the data randomization test.*

4.2. Ablation Tests

Fig. 2 and Fig. 15 (in Appendix C) show the results of the ablation tests (high-drop and low-drop, respectively, for 500 randomly sampled individual neighborhoods across all baselines with and without absolute-value conversion). Overall, we observe a more expected behavior when using the absolute values of the obtained attributions, as in these cases the average class probability goes down

when ablating high-relevance points, with the average class probability staying stable or even increasing when ablating low-relevance points. We observe different results across the evaluated baselines, with the zero-vector baseline showing a stronger effect.

Fig. 3 and Fig. 16 (in Appendix C) show the results of the high-drop and low-drop ablation tests for the full 3D point cloud, using SCHWEGLER, MÜLLER, and REITERER [SMR23] overwriting method and our averaging method to compute the final attribution scores used for ablation and saliency map generation. We observe a similar pattern as with individual neighborhoods (Figs. 2 and 15), with the absolute values of the attributions displaying the expected behavior (i.e., the average class probability being reduced when ablating high-attribution points, and vice versa). Consistent with the individual neighborhood ablation tests, we observe a different behavior across the evaluated baselines; the zero-vector baseline combined with our averaging method shows the best overall results. Surprisingly, when preserving the signs of the attribution values, the average class probability increases as larger percentages of high-attribution points are removed for the individual neighborhoods and the aggregated results. Furthermore, when ablating low-attribution points, we see a stronger tendency for the average class probability to stay stable or increase when using absolute values. This indicates that high-attribution points might be incorrectly assigned negative signs during the IG computation.

4.3. Saliency Map Visualization

Since the zero-vector baseline shows the best overall results in the ablation tests, in most cases showing a larger decrease in the average class probability when removing high-attribution points (Figs. 2 and 3) and vice versa (Figs. 15 and 16), we choose it to generate the final saliency maps. Figs. 4 and 5 and Figs. 8 to 11 in the appendix show a comparison between the outputs from our proposed averaging approach and the overwriting approach from [SMR23]. Our method produces saliency maps highlighting more clearly the points belonging to the *Roadline* class, and points close to them. For the *Building* class, both approaches show similar results, with the divergent color map showing a higher number of positively attributed points within the building structure. Our approach highlights a large number of points for the *Vegetation* class, with many belonging to surrounding structures. However, this result is likely due to the attribution capping threshold used (see Appendix B). Finally, Fig. 6 shows saliency maps for all classes and a zero-vector baseline using our averaging approach.

5. Discussion and Conclusions

The model parameter and data randomization tests show consistent results across all baselines and evaluated classes with low Spearman correlation scores, indicating no evidence that IG is insensitive to the model training and data labeling process, pointing towards the usefulness of IG as a tool for debugging DL models for 3D point cloud semantic segmentation.

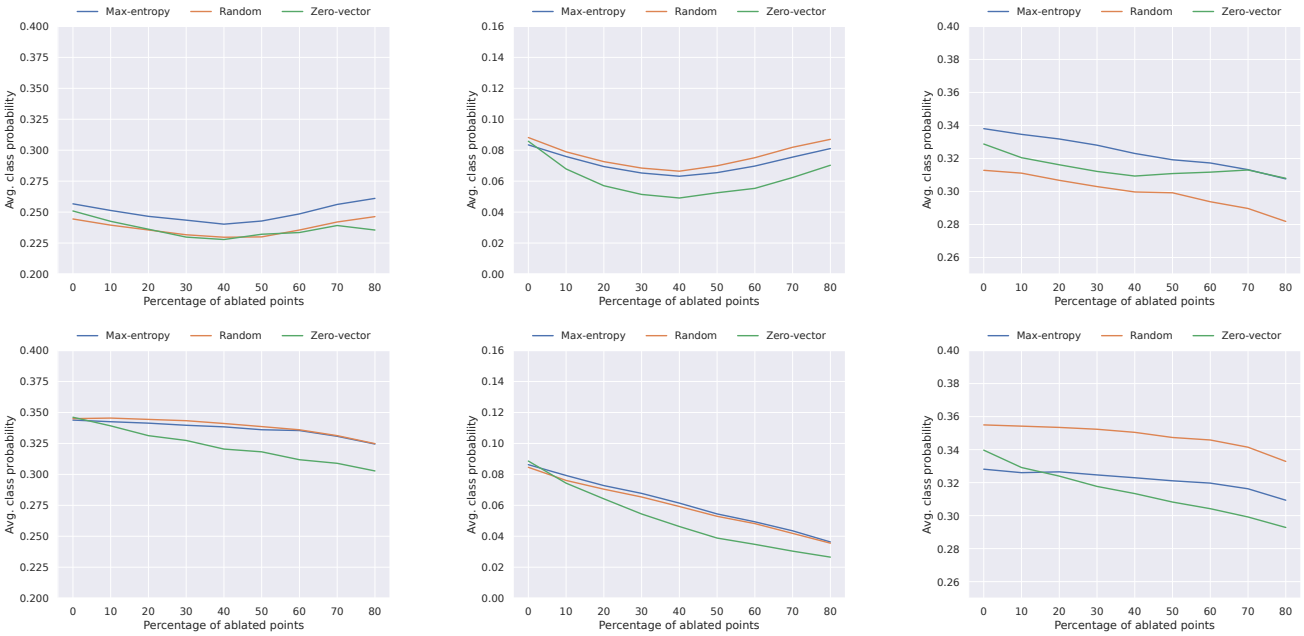
However, the IG method shows instability with respect to the choice of baseline, as the results change between baselines for the evaluated classes, both in terms of the saliency maps themselves and the ablation test results. Overall, the zero-vector baseline produces clearer saliency maps and performs better in the ablation studies compared to the random and max-entropy baselines for the evaluated DL model and semantic segmentation task.

The ablation test results are better when using the absolute values of the attributions instead of the signed values. This result suggests that the IG method with the zero-vector baseline can correctly identify points that have a high overall influence on the prediction probability. Nevertheless, when ablating high-contribution points using the original signs of the obtained attributions, we see an unexpected increase in the average class probabilities for the individual neighborhoods and the aggregated saliency maps. This behavior indicates that IG fails to properly distinguish between points with positive and negative contributions to the prediction probability. A limitation of the ablation test results is that the observed changes might be because the ablated model inputs no longer fit into the distribution of training data learned by the model, rather than to the effect of removing high- or low-attribution points [HEKK19]. In our study, this is mitigated in the case of the aggregated saliency maps, since we directly remove points from the 3D point cloud, reducing the need for the model to extrapolate on the modified inputs [HMZ21]. Furthermore, we observe a consistent behavior in both the per-neighborhood and aggregated ablation studies. Nevertheless, it is necessary to evaluate the IG across a wider variety of DL architectures and object classes in future works, as it is possible that these results would differ in such cases.

Regarding the aggregation of the attributions output by the IG method, we find that averaging them for points that received multiple prediction scores and attribution values produces sharper saliency maps and yields better results in the ablation tests. Intuitively, points that consistently receive high attribution scores across multiple overlapping neighborhoods should contribute to a higher degree to the model output. Nevertheless, exploring different aggregation approaches that address the inherent loss of detail stemming from the aggregation process stands as a promising avenue for further research.

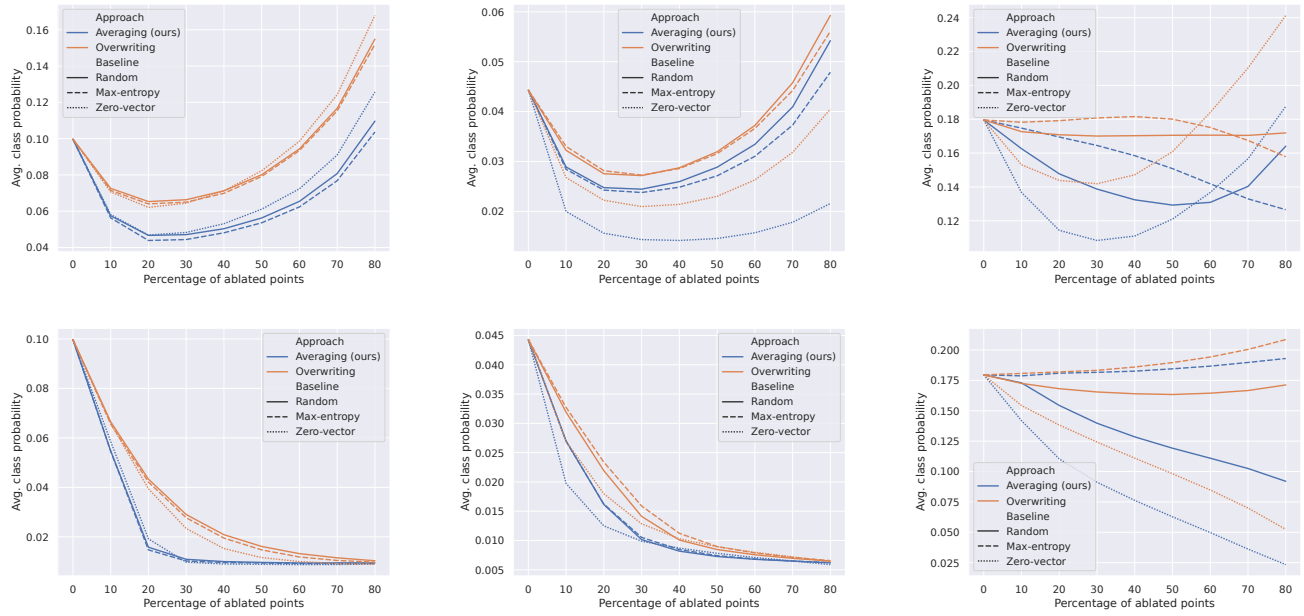
Furthermore, we observe a change in the distribution of the attribution values obtained by our averaging approach (see Appendix B), which in turn influences decisions related to the visualization of the saliency maps (e.g., different capping thresholds for the *Vegetation* class, affecting its corresponding saliency map). However, given that we see high-attribution points for this class in objects unrelated to the class in question (e.g., a large number of high-attribution points in the road surface), these results may reflect that the model has learned an unintended task (i.e., a task different from vegetation segmentation) by modeling the loss function or data in a way that allows for an unintended abstraction - which could also explain the inconsistent behavior observed in the ablation test results for this class.

Our data processing pipeline is constrained to generate inputs of fixed size for the DL models. This constraint allows us to efficiently implement the IG method because only one baseline needs to be generated for the whole process. However, when using data processing pipelines and DL architectures that work with variable input sizes, scalability issues could arise for IG since this would require the generation of different baselines for each input size.



(a) *Building class.* **(b)** *Roadline class.* **(c)** *Vegetation class.*

Figure 2: *Predicted probability scores after removing high-attribution points (lower is better) for different classes and baselines, when using signed attribution values (first row) or absolute attribution values (second row), averaged over 500 randomly sampled subsections (neighborhoods).*



(a) *Building class.* **(b)** *Roadline class.* **(c)** *Vegetation class.*

Figure 3: *Predicted probability scores after removing high-attribution points (lower is better) for different classes and all baselines, when using signed attribution values (first row) or absolute attribution values (second row) for the final saliency maps using our averaging approach and the overwriting approach from [SMR23].*

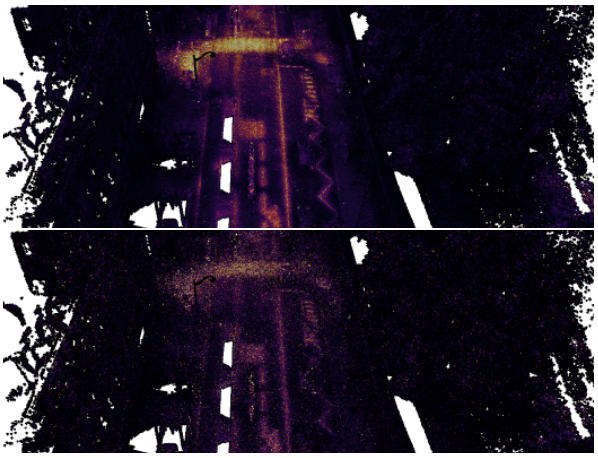


Figure 4: *Saliency maps for the Roadline class with a zero-vector baseline. Absolute-value visualization with our averaging approach (top) and the overwriting approach from [SMR23] (bottom).*

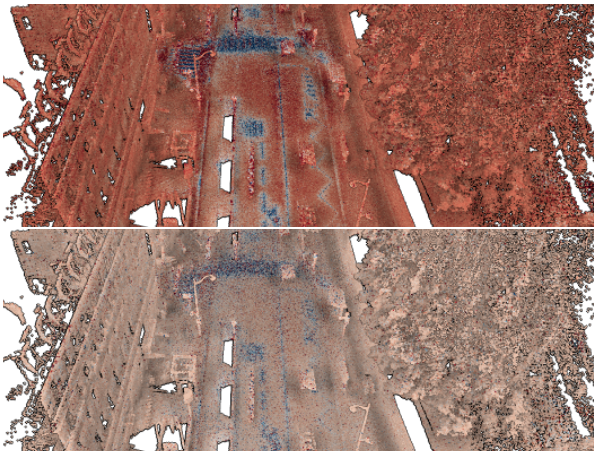


Figure 5: *Saliency maps for the Roadline class with a zero-vector baseline. Diverging visualization with our averaging approach (top) and the overwriting approach from [SMR23] (bottom).*

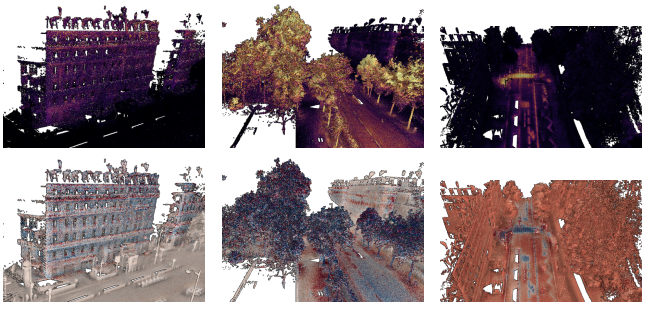


Figure 6: *Saliency maps for all evaluated classes using a zero-vector baseline, with absolute-value (top) and diverging (bottom) visualizations.*

Generating multiple baselines would add a significant amount of computational load, especially when using approaches such as the max-entropy baseline, which requires a gradient ascent process for each sample to be explained.

Although through visual inspection, the absolute-value saliency maps highlight areas that are intuitive from a human perspective (e.g., high attribution scores on building structures and road marks), the divergent saliency maps show a mix of highly positive and highly negative attributions in such regions, following no obvious pattern. This behavior could be due to the model itself learning unintended patterns; however, it could also be due to visualization artifacts stemming from decisions taken for the saliency map visualization (e.g., defining capping thresholds in too low or too high values). XAI visualization techniques better suited to produce human-understandable explanations stands as a promising avenue for future work. However, such approaches should take care to accurately reflect the models' behavior, avoiding confirmation biases and reducing overreliance and false expectations on the performance of the models [VJG*23].

Our results suggest that the IG method can identify points with a high contribution towards a model output for 3D point cloud semantic segmentation, and it is sensitive to model parameters and training data, showing its potential as a tool for use cases such as model debugging. Nevertheless, care should be taken to select an appropriate baseline, as this can significantly impact the results. Furthermore, our experiments indicate that IG struggles to correctly distinguish between positive and negative impact of points towards a model output, which should be taken into consideration when interpreting the generated explanations.

Acknowledgements

This work was partially funded through grants by the Service-Oriented Systems Engineering Research School of the Hasso Plattner Institute and by the Federal Ministry of Education and Research, Germany through grant 033L305 ("TreeDigitalTwins") and grant 01IS22062 ("AI research group FFS-AI"). We thank the anonymous reviewers for their valuable feedback.

References

[AGM*18] ADEBAYO, JULIUS, GILMER, JUSTIN, MUELLY, MICHAEL, et al. "Sanity Checks for Saliency Maps". *Advances in Neural Information Processing Systems*. Ed. by BENGIO, S., WALLACH, H., LAROCHELLE, H., et al. Vol. 31. Curran Associates, Inc., 2018 1–4.

[DDR*21] DESCHAUD, JEAN-EMMANUEL, DUQUE, DAVID, RICHA, JEAN PIERRE, et al. "Paris-CARLA-3D: A Real and Synthetic Outdoor Point Cloud Dataset for Challenging Tasks in 3D Mapping". *Remote Sensing* 13.22 (2021) 2, 3.

[GAZ19] GHORBANI, AMIRATA, ABID, ABUBAKAR, and ZOU, JAMES. "Interpretation of Neural Networks Is Fragile". *AAAI Conference on Artificial Intelligence* 33.01 (2019), 3681–3688 3.

[GHD*17] GÉLARD, WILLIAM, HERBULOT, ARIANE, DEVY, MICHEL, et al. "Leaves Segmentation in 3D Point Cloud". *Advanced Concepts for Intelligent Vision Systems*. Ed. by BLANC-TALON, JACQUES, PENNE, RUDI, PHILIPS, WILFRIED, et al. Cham: Springer International Publishing, 2017, 664–674 1.

[GWY20] GUPTA, ANANYA, WATSON, SIMON, and YIN, HUJUN. "3D Point Cloud Feature Explanations Using Gradient-Based Methods". *International Joint Conference on Neural Networks*. 2020, 1–8 1, 2.

[HDD*92] HOPPE, HUGUES, DEROSE, TONY, DUCHAMP, TOM, et al. "Surface Reconstruction from Unorganized Points". *Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '92. New York, NY, USA: Association for Computing Machinery, 1992, 71–78 4.

[HEKK19] HOOKER, SARA, ERHAN, DUMITRU, KINDERMANS, PIETER-JAN, and KIM, BEEN. "A Benchmark for Interpretability Methods in Deep Neural Networks". *Advances in Neural Information Processing Systems*. Ed. by WALLACH, H., LAROCHELLE, H., BEYGELZIMER, A., et al. Vol. 32. Curran Associates, Inc., 2019 6.

[HHH22] HOLMBERG, LARS, HELGSTRAND, CARL JOHAN, and HULTIN, NIKLAS. "More Sanity Checks for Saliency Maps". *Foundations of Intelligent Systems*. Ed. by CECI, MICHELANGELO, FLESCA, SERGIO, MASCIARI, ELIO, et al. Cham: Springer International Publishing, 2022, 175–184 3.

[HMZ21] HOOKER, GILES, MENTCH, LUCAS, and ZHOU, SIYU. "Unrestricted Permutation Forces Extrapolation: Variable Importance Requires at Least One More Model, or there is no Free Variable Importance". *Statistics and Computing* 31.6 (2021), 82 6.

[HSM*22] HOLZINGER, ANDREAS, SARANTI, ANNA, MOLNAR, CHRISTOPH, et al. "Explainable AI Methods - A Brief Overview". *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020*. Ed. by HOLZINGER, ANDREAS, GOEBEL, RANDY, FONG, RUTH, et al. Cham: Springer International Publishing, 2022, 13–38 2, 3.

[HWLH24] HEDSTRÖM, ANNA, WEBER, LEANDER, LAPUSCHKIN, SEBASTIAN, and HÖHNE, MARINA. "A Fresh Look at Sanity Checks for Saliency Maps". *Explainable Artificial Intelligence*. Ed. by LONGO, LUCA, LAPUSCHKIN, SEBASTIAN, and SEIFERT, CHRISTIN. Cham: Springer Nature Switzerland, 2024, 403–420 3.

[HYX*20] HU, QINGYONG, YANG, BO, XIE, LINHAI, et al. "RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds". *IEEE Computer Vision and Pattern Recognition Conference* (2020) 2, 3, 10.

[KAWH19] KUMAR, A., ANDERS, K., WINIWARTER, L., and HÖFLE, B. "Feature Relevance Analysis For 3D Point Cloud Classification Using Deep Learning". *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W5 (2019), 373–380 4.

[KHA*19] KINDERMANS, PIETER-JAN, HOOKER, SARA, ADEBAYO, JULIUS, et al. "The (Un)reliability of Saliency Methods". *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by SAMEK, WOJCIECH, MONTAVON, GRÉGOIRE, VEDALDI, ANDREA, et al. Cham: Springer International Publishing, 2019, 267–280 3.

[KK24] KURIYAL, ABHISHEK and KUMAR, VAIBHAV. *Towards Explainable LiDAR Point Cloud Semantic Segmentation via Gradient Based Target Localization*. 2024 2, 3.

[LBB*23] LA ROSA, B., BLASILLI, G., BOURQUI, R., et al. "State of the Art of Visual Analytics for eXplainable Deep Learning". *Computer Graphics Forum* 42.1 (2023), 319–355 1.

[mat24] MATPLOTLIB. *Choosing Colormaps in Matplotlib*. Accessed: 2024-06-13. 2024 4.

[MPF*22] MATRONE, FRANCESCA, PAOLANTI, MARINA, FELICETTI, ANDREA, et al. "BubblEX: An Explainable Deep Learning Framework for Point-Cloud Classification". *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), 6571–6587 2.

[PDS18] PETSIUK, VITALI, DAS, ABIR, and SAENKO, KATE. *RISE: Randomized Input Sampling for Explanation of Black-box Models*. 2018 2, 3, 5.

[PGK02] PAULY, M., GROSS, M., and KOBBELT, L.P. "Efficient Simplification of Point-Sampled Surfaces". *IEEE Visualization*. 2002, 163–170 4.

[QSMG17] QI, CHARLES R., SU, HAO, MO, KAICHUN, and GUIBAS, LEONIDAS J. "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation". *IEEE Computer Vision and Pattern Recognition Conference*. IEEE, 2017, 77–85 2.

[SCD*17] SELVARAJU, RAMPRASAATH R., COGSWELL, MICHAEL, DAS, ABHISHEK, et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". *International Conference on Computer Vision*. 2017, 618–626 2.

[SDBR15] SPRINGENBERG, J.T., DOSOVITSKIY, A., BROX, T., and RIEDMILLER, M. "Striving for Simplicity: The All Convolutional Net". *International Conference on Learning Representations (workshop track)*. 2015 2.

[SLL20] STURMFELS, PASCAL, LUNDBERG, SCOTT, and LEE, SU-IN. "Visualizing the Impact of Feature Attribution Baselines". *Distill* (2020). https://distill.pub/2020/attribution-baselines 3, 5.

[SMR23] SCHWEGLER, MARKUS, MÜLLER, CHRISTOPH, and REITERER, ALEXANDER. "Integrated Gradients for Feature Assessment in Point Cloud-Based Data Sets". *Algorithms* 16.7 (2023) 1–4, 6–8, 10, 11, 13.

[STK*17] SMILKOV, DANIEL, THORAT, NIKHIL, KIM, BEEN, et al. *SmoothGrad: Removing Noise by Adding Noise*. 2017 4.

[STY17] SUNDARARAJAN, MUKUND, TALY, ANKUR, and YAN, QIQI. "Axiomatic Attribution for Deep Networks". *International Conference on Machine Learning*. Ed. by PRECUP, DOINA and TEH, YEE WHYE. Vol. 70. Proceedings of Machine Learning Research. 2017, 3319–3328 2–5.

[SVZ14] SIMONYAN, KAREN, VEDALDI, ANDREA, and ZISSERMAN, ANDREW. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014 2.

[Tan23a] TAN, HANXIAO. "Maximum Entropy Baseline for Integrated Gradients". *International Joint Conference on Neural Networks*. 2023, 1–8 3, 5.

[Tan23b] TAN, HANXIAO. "Visualizing Global Explanations of Point Cloud DNNs". *Winter Conference on Applications of Computer Vision*. 2023, 4730–4739 2.

[TK22] TAN, HANXIAO and KOTTHAUS, HELENA. "Surrogate Model-Based Explainability Methods for Point Cloud NNs". *Winter Conference on Applications of Computer Vision*. 2022, 2239–2248 2.

[TK23] TAN, HANXIAO and KOTTHAUS, HELENA. "Explainability-Aware One Point Attack for Point Cloud Neural Networks". *Winter Conference on Applications of Computer Vision*. 2023, 4570–4579 2.

[TQD*19] THOMAS, HUGUES, QI, CHARLES R., DESCHAUD, JEAN-EMMANUEL, et al. "KPConv: Flexible and Deformable Convolution for Point Clouds". *International Conference on Computer Vision* (Seoul, Korea). IEEE, 2019, 6410–6419 4.

[TSS23] TAYYUB, JAWAD, SARMAD, MUHAMMAD, and SCHÖNBORN, NICOLAS. "Explaining Deep Neural Networks for Point Clouds Using Gradient-Based Visualisations". *Asian Conference on Computer Vision*. Macao, China: Springer-Verlag, 2023, 155–170 1–3, 5.

[VJG*23] VASCONCELOS, HELENA, JÖRKE, MATTHEW, GRUNDE-MCLAUGHLIN, MADELEINE, et al. "Explanations Can Reduce Overreliance on AI Systems During Decision-Making". *Proceedings ACM Human-Computational Interaction* 7.CSCW1 (2023) 8.

[WK19] WANG, QIAN and KIM, MIN-KOO. "Applications of 3D Point Cloud Data in the Construction Industry: A Fifteen-Year Review from 2004 to 2018". *Advanced Engineering Informatics* 39 (2019), 306–319 1.

[ZCY*19] ZHENG, TIANHANG, CHEN, CHANGYOU, YUAN, JUNSONG, et al. "PointCloud Saliency Maps". *International Conference on Computer Vision*. 2019, 1598–1606 1–3.

[ZWQF20] ZIWEN, CHEN, WU, WENXUAN, QI, ZHONGANG, and FUXIN, LI. "Visualizing Point Cloud Classifiers by Curvature Smoothing". *British Machine Vision Virtual Conference* (2020) 1, 2.