S1. Supplementary materials

These materials are supplements for the main paper. In section S1.1 we present the best hyper-parameter settings of our explainability method. In section S1.2, we visually compare the proposed flip operation with the most popular ones currently available.

S1.1. Hyper-parameter configuration

S1.1.1 Number of clusters & samples

As the majority of point cloud-based classification algorithms incorporate sampling operations in their preprocessing, with 1024 generally being the most common input size, in this section we first investigate the most appropriate hyper-parameter setting for the input size 1024.

We sample the number of clusters with 20, 64, 128 and 1024. To ensure that 10 sets of prediction scores are obtained while at most 50% of the points are flipped in the proposed model-independent verification method, we set the number of points per cluster to approximate 5% of the input size as the minimum value, i.e., 20 clusters. As the other extreme, each point is individually regarded as a cluster, where the number of clusters is 1024. 64 and 128 are randomly picked considering the processing time. We selected 10^2,$10^{2.5}$,10^3,$10^{3.5}$ and 10^4 as the incremental number of perturbation samples. Tables S2 and S3 present the local fidelities with various hyper-parameter settings. Note that $\hat{R^2_\omega}$ is more convincing when comparing different S because R^2_ω is sensitive to the number of samples S. When comparing different C, both R^2_ω and $\hat{R^2_\omega}$ can be considered as references. But since R2 is significant only for $S > C$, R^2_ω is chosen as a more complete observation. The results show that the proximity benefit from both greater C and S, while there is little utility but severe time-consuming when S exceeds a certain threshold (S magnifies from 10^3 to 10^4). On the other hand, augmenting C dramatically enhances the fidelity of the surrogate model due to the alleviation of attribution neutralization. However, explaining points individually suffers from loss of the semantic meaning of clusters (see figure 1), which is more incomprehensible to humans.

The plausibilities under different combinations of hyper-parameters are depicted in figures S2 and S3. Identical to the local fidelity, the plausibility benefits from enlarging the number of clusters as well, whereas they both suffer from the semantic issue. On the other hand, enhancing the plausibility via additional perturbation samples is ineffective. At the expense of 10 times the number of perturbed samples, only less than 0.1 additional plausibility is achieved, which is unacceptable for point clouds with the high demand of real-time. The time costs under a various number of samples are presented in table S1. Since the run time of the explaining process is independent of the number of clusters, we only demonstrate its relationship to the number of perturbation samples S.

Though point-wise explanation outperforms the cluster-based one, it assumes a sufficiently large number of perturbation samples. However in industrial applications, models may involve large-scale point clouds as inputs that are composed of millions of points. We therefore strongly recommend an appropriate number of clusters in the trade-off between performance and real-time capability. For instance, a point-wise explanation of an instance composed of 10^4 points with 10^4 samples takes 507 seconds, and this cost explosively grows to 3909 seconds when perturbing 10^5 number of samples. Note that a theoretically complete explanation for the aforementioned instance requires 2^{10^4} perturbed samples, whose processing time is apparently unacceptable in practice.

To summarize, we find that setting $C = 128$ and $S = 10^3$ is most suitable for the most popular point cloud classification model with 1024 sampling points as input. Although the point-by-point explanation dominates in performance, when dealing with large-scale point clouds, a suitable choice of cluster number for explanation is recommended considering the time cost.

S1.1.2 Sensitivity study of kernel width

The kernel width w is another relevant hyper-parameter that impacts the explanations. To explore the best performing kernel width we sample the kernel width from 0.05 to 0.3 with step 0.05, which is empirically an appropriate range of kernel width for point clouds, and with the remaining hyper-parameters identical. The results of the sensitivity study are presented in figure S1. According to the results, we observe that the performance peaks at a kernel width of 0.1 and then inconspicuous degrades with further increasing kernel width. Our proposed point cloud-applicable explainability method exhibits low sensitivity to the kernel width in an appropriate interval.

S	10^2	$10^{2.5}$	10^3	$10^{3.5}$	10^4
Time(s)	1.05	2.21	5.62	17.51	58.96

Table S1. Average processing time (in seconds) of LIME on single point cloud instance concerning the number of perturbed samples. All values are recorded as the average of 1000 experiments.

S1.2. VISF VS. traditional flipping operations

Fig. S4 compares the three aforementioned flipping operations in section 3.1.3. As can be seen from the figure, the currently most popular flipping operations (i.e. replacing the candidate points to be flipped with zeros and means of the rest points) fail to eliminate additional interference of flipping. As there is a lumpy collection of overlapping points at a specific location, it is difficult to determine whether the variations in predicted scores are independent

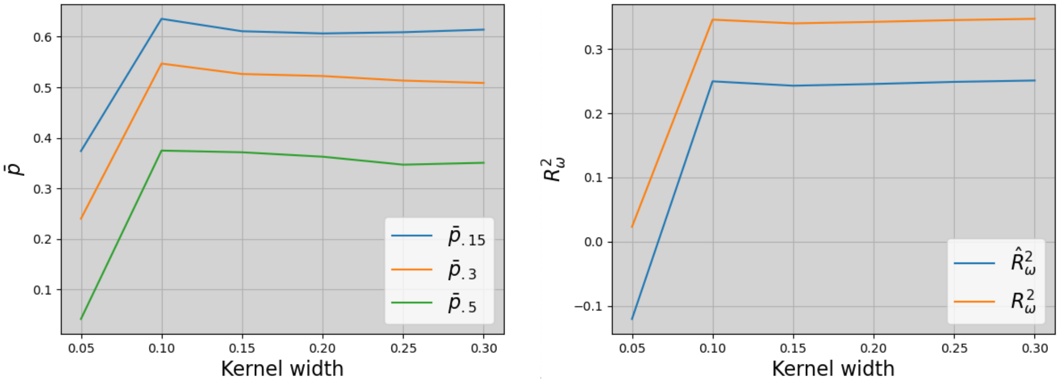

Figure S1. Plausibility (left) and local fidelity (right) of proposed method under different kernel widths w. For conciseness, we only demonstrate the coefficient of determination R^2_ω as the distance metrics (e.g. L_1) share the similar trend with R^2_ω.

C	L_m	L_1	L_1^ω	L_2	L_2^ω	R^2_ω	$\hat{R}^2_\omega$
20	4.28×10^{-1}	7.84×10^{-1}	2.91×10^{-1}	3.02	4.23×10^{-1}	0.239	0.223
64	4.47×10^{-2}	1.88×10^{-1}	1.27×10^{-1}	2.77×10^{-1}	1.21×10^{-1}	0.249	0.198
128	1.03×10^{-2}	8.90×10^{-2}	6.95×10^{-2}	7.84×10^{-2}	4.82×10^{-2}	0.345	0.249
1024	$\mathbf{1.28 \times 10^{-4}}$	$\mathbf{1.00 \times 10^{-2}}$	$\mathbf{8.73 \times 10^{-3}}$	$\mathbf{1.73 \times 10^{-3}}$	$\mathbf{1.41 \times 10^{-3}}$	**0.883**	$\backslash$

Table S2. Local fidelity metrics of different C with 1000 perturbation samples.

S	L_m	L_1	L_1^ω	L_2	L_2^ω	R^2_ω	$\hat{R}^2_\omega$
10^2	5.30×10^{-2}	$\mathbf{1.86 \times 10^{-1}}$	$\mathbf{7.78 \times 10^{-2}}$	3.03×10^{-1}	7.59×10^{-2}	**0.389**	-0.728
10^3	4.47×10^{-2}	1.88×10^{-1}	1.27×10^{-1}	2.77×10^{-1}	$\mathbf{1.21 \times 10^{-1}}$	0.249	0.198
10^4	$\mathbf{2.64 \times 10^{-2}}$	1.88×10^{-1}	1.38×10^{-1}	$\mathbf{2.63 \times 10^{-1}}$	1.34×10^{-1}	0.209	**0.204**

Table S3. Local fidelity metrics of different S with 64 clusters.

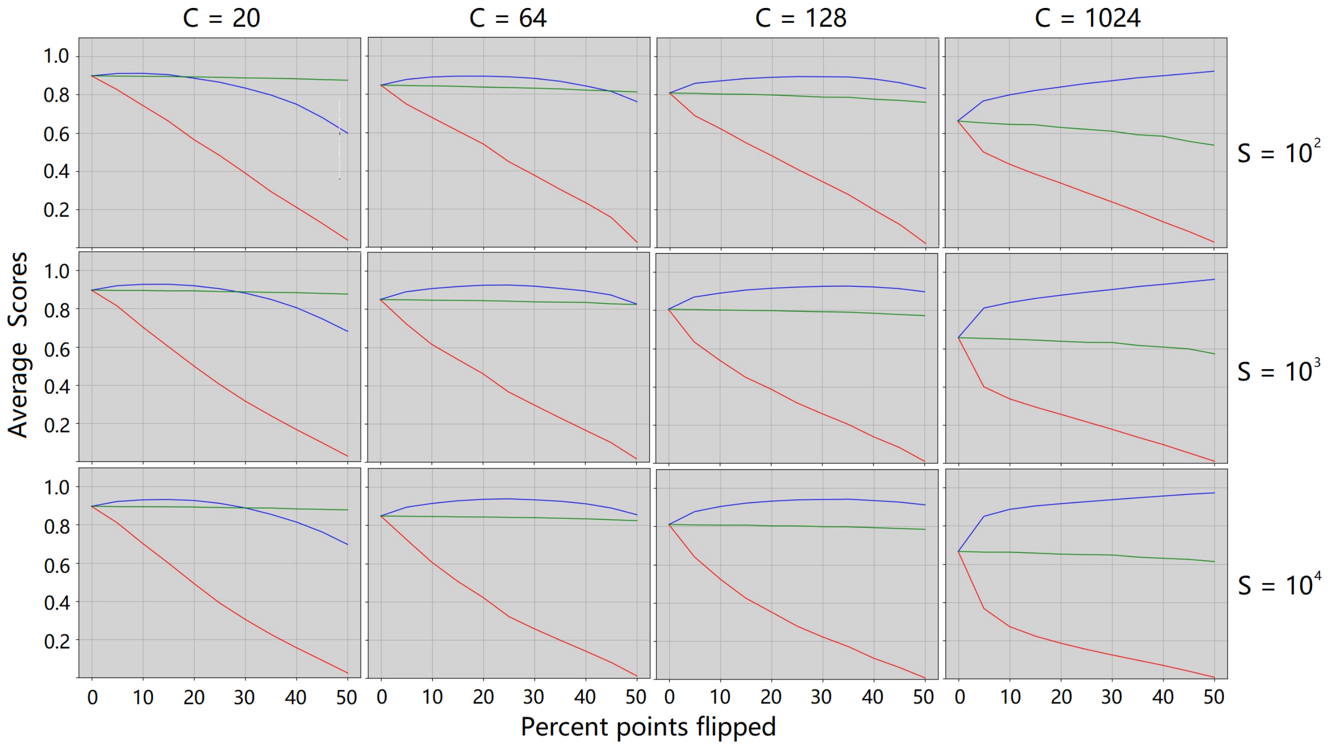

Figure S2. Means of prediction scores for different combinations of parameters of LIME. C denotes the number of clusters (features) and S denotes the number of samples used to train the surrogate model. The red and blue lines indicate the means of flipping positive and negative contributing points, while the green line indicates random flipping of the same percentage of arbitrary points.

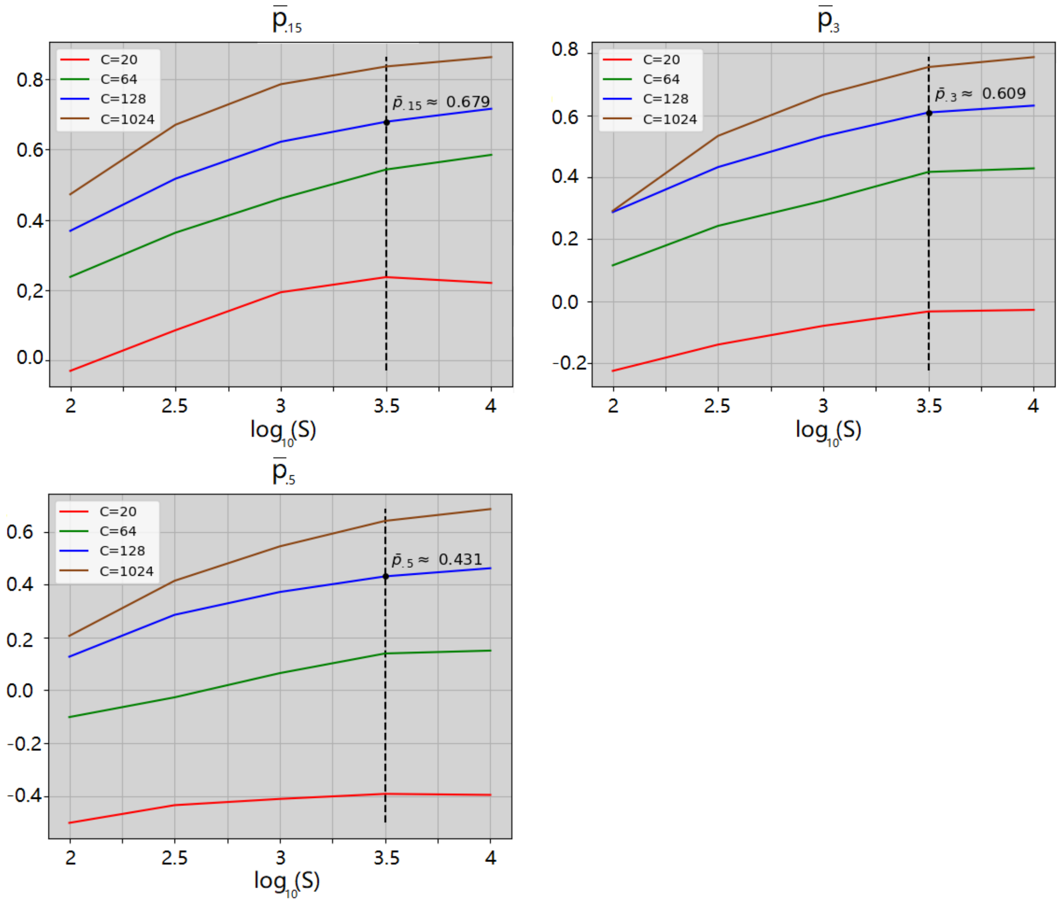

Figure S3. Plausibility with different hyper-parameter settings, i.e. $\bar{p}_{.15}$, $\bar{p}_{.3}$ and $\bar{p}_{.5}$.

of this lump of points. In contrast, the instances flipped by VISF contain no similar lump, leading to a more convincing association of predicted scores with important features.

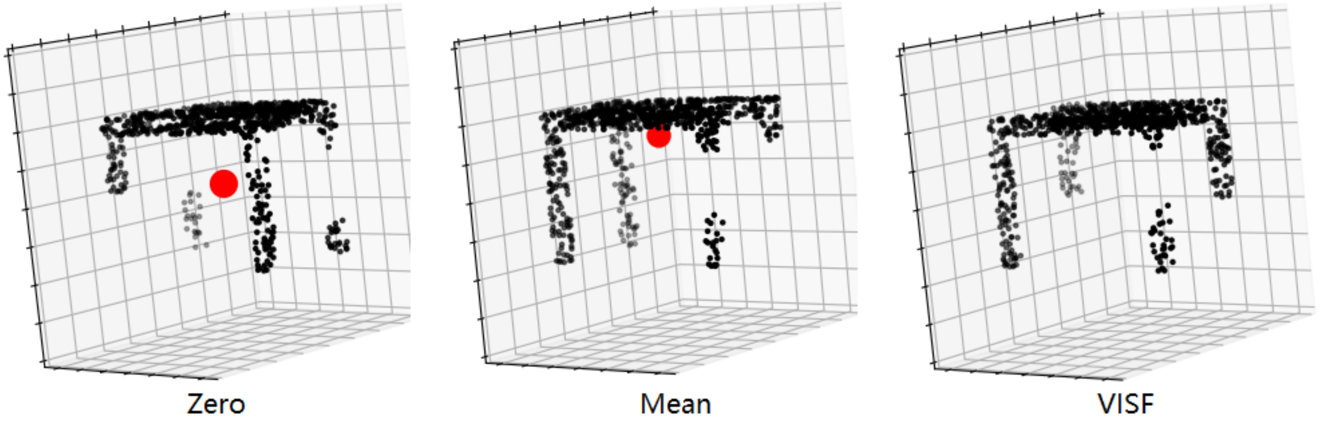

Figure S4. Visualization of three different flipping methods (20% points are flipped). Zero and mean denotes replacing the candidate points to be flipped with zeros and the mean of the three axes of rest points respectively. VISF denotes our Variable input size flipping. The more overlapped points located at the same coordinates, the larger the diameter of that point in the image. For better observation, the points of the flipped destination are marked in red.