Explainable Artificial Intelligence (XAI) for Point Cloud Perception in Autonomous Driving: A survey

Aitor Iglesias[1,2], Ruben Naranjo[1,2], Nerea Aranjuelo[1], Ignacio Arganda-Carreras[2,3,4,5], and Marcos Nieto[1]

[1] Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, Spain
[2] University of the Basque Country (UPV/EHU), Donostia - San Sebastián, Spain
[3] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[4] Donostia International Physics Center (DIPC), Donostia - San Sebastián, Spain
[5] Biofisika Institute, Leioa, Spain

Abstract. This survey provides the first comprehensive review of Explainable Artificial Intelligence (XAI) methods applied to point cloud-based perception in autonomous driving. Although Deep Learning has significantly advanced 3D perception tasks such as object detection and semantic segmentation, the opacity of these models is a challenges for safety-critical systems. To address this, the paper examines existing XAI for point cloud approaches and categorizes them based on a previously defined XAI-specific taxonomy. The review highlights the challenges and limitations of the current state of the art, which includes the lack of standardized evaluation metrics and the limited number of XAI methods tailored specifically to point cloud data in the field of autonomous driving. In this work, we also outline future research directions, emphasizing the need for point cloud–native XAI techniques, stronger validation frameworks, and the integration of emerging vision-language models to enhance explainability and reliability in autonomous driving.

Keywords: Explainable AI · Artificial intelligence · Survey · Point cloud · Computer Vision · Autonomous Driving

1 Introduction

Autonomous driving (AD) aims to enhance transportation safety, efficiency, and accessibility. To operate safely, autonomous vehicles must perceive and understand their surroundings in real time. This perception relies on sensors such as cameras, radar, and LiDAR, which together form a comprehensive view of the environment. Among these sensors, LiDAR measures distances using laser pulses, creating 3D point clouds that capture the scene's geometry. Unlike cameras, it provides precise depth and spatial data, making it vital for object detection, localization, tracking, and 3D mapping.

AD relies on Deep Learning (DL) methods for understanding its environment. Although DL has greatly advanced perception performance, its decision processes often remain opaque. Understanding the reasoning behind model outputs is crucial for verifying reliability and ensuring safe deployment. Explainable Artificial Intelligence (XAI) addresses this issue by developing methods that inform humans of the reason behind decisions made by AI. Explainability is particularly relevant in AD, where decisions can have direct safety implications.

While multiple surveys have examined point cloud–based perception in autonomous driving [15,34], they do not address XAI. Recently, XAI in autonomous driving research has grown considerably, with many surveys [12, 19], including also multimodal sensor fusion [39], and ontology-based data models [23]. Yet, these reviews rarely tackle explainability specifically for point cloud data. One survey has explored XAI methods for point clouds [21], primarily in classification tasks, but none have analyzed these methods from the perspective of autonomous driving.

To address this gap, our paper studies the unexplored intersection of XAI, point cloud perception, and autonomous driving. Our main contribution is compiling and organizing the literature on explainability methods for LiDAR-based perception models, categorizing the methods based on previously established XAI taxonomy. In addition, we summarize the evaluation techniques used in these studies and identify current limitations and open research directions in this emerging field.

2 Explainable Artificial Intelligence (XAI)

To provide a structured and systematic overview of the existing literature, we adopt the taxonomy introduced by Ali et al. [2], which categorizes explainability techniques according to their methodological contribution. This taxonomy distinguishes between two primary groups: Model Explainability and Post-hoc Explainability.

Model explainability focuses on developing models that are inherently transparent, allowing researchers to directly inspect their internal structure and training procedures in order to understand the decision-making process. Within this group, the taxonomy distinguishes several categories:

- **Inherently Interpretable Models**: This group encompasses "white-box" models, such as linear regression and decision trees, which provide algorithmic transparency by design.
- **Hybrid Explainable Models**: These architectures aim to achieve both high accuracy and interpretability by combining the expressive power of complex "black-box" models (like Deep Neural Networks) with the transparent structure of inherently interpretable techniques. Examples include the Deep k-Nearest Neighbors (DkNN) algorithm, which performs kNN inference on the hidden representations extracted by a neural network and Self-Explaining Neural Networks (SENN), which generalize linear classifiers by using networks to learn features and coefficients.

- **Joint Prediction and Explanation**: These models are explicitly architected and trained to output a target prediction alongside a corresponding explanation simultaneously. Frameworks like Teaching Explanations for Decisions (TED) augment the training dataset with domain-expert rationales, forcing the model to learn and decode both the output label and the reasoning behind it. Other approaches within this category, such as Rationalizing Neural Predictions (RNP), utilize a generator-encoder structure to extract brief, cohesive input text fragments that act as the rationale for the prediction.
- **Explainability through Architectural Adjustments**: This technique enhances transparency by introducing structural modifications to a standard black-box architecture. For instance, adding a Prototypes Layer (PL) between convolutional and fully connected layers allows the network to classify images by comparing different parts of an input to a learned collection of prototypical components. Similarly, attention mechanisms highlight the most relevant parts of the input data for a specific task.
- **Explainability through Regularization**: This methodology improves the intrinsic explainability of complex models by incorporating specific penalty terms into the loss function during the training phase. For example, Tree Regularization encourages the neural network to learn decision boundaries that can be accurately approximated by a simple decision tree. Other approaches penalize input gradients that contradict established domain knowledge stored as binary mask matrices, ensuring the model makes predictions for the right reasons.

Post-hoc explainability refers to techniques applied after a model has been trained, with the objective of interpreting and understanding the decision-making process of "black-box" models.

- **Attribution Methods**: These methods assign a relevance score or contribution value to each individual input feature (such as an image pixel or a word) to quantify its influence on a specific model output. This category relies heavily on techniques like backpropagation (e.g., computing gradients of the output with respect to the input) and perturbation (measuring output variation after masking or altering inputs).
- **Visualization Methods**: These techniques provide graphical representations of the underlying patterns learned by supervised models. Partial Dependence Plots (PDP) illustrate the marginal, global effect that one or two features have on the average predictions of a model. Accumulated Local Effects (ALE) offer an unbiased alternative to PDPs by averaging and accumulating local prediction changes, making them robust when input features are highly correlated.
- **Example-based Explanation Methods**: Also known as case-based explanations, these methods elucidate the model's behavior by referring to specific, illustrative data instances. This category encompasses techniques such as counterfactuals, which are often synthetically generated "contrary-to-fact" examples that demonstrate the minimal necessary perturbations or

conditions required to reverse or alter the model's decision to a desired alternative outcome. Another approach involves adversarial examples, while traditionally used to evaluate model robustness and vulnerabilities, can also be leveraged to generate analogical explanations by comparing familiar and unfamiliar domains, or contrastive explanations that provide competing evidence to clarify complex decision boundaries.

- **Game Theory Methods**: Inspired by game theory, these methods calculate the exact contribution of each feature to a prediction. The model's prediction is treated as the game's gain, and individual features are treated as "players". Shapley Values compute the weighted average of a feature's marginal contribution across all possible feature combinations. SHAP (SHapley Additive exPlanations) provides a unified additive feature attribution framework that estimates these values to deliver both local and global interpretations.
- **Knowledge Extraction Methods**: This approach seeks to retrieve the complex, hidden knowledge learned by Neural Networks and translate it into an interpretable format. Rule Extraction produces a readable approximation of the network's behavior using symbolic representations, such as IF-THEN or M-of-N logical rules. This can be done via decompositional methods (analyzing individual neurons). Another example is Model Distillation which involves transferring knowledge from a black-box teacher model into a simpler, naturally interpretable student model.
- **Neural Methods**: These techniques are specifically tailored to analyze the internal states and representations of neural networks. Influence Methods evaluate feature significance by analyzing internal parameters or errors. Whereas Concept Methods, such as Concept Activation Vectors (CAVs), move beyond raw input features (like pixels) to explain the internal states of neural networks using human-understandable concepts by analyzing layer activations.

Although this is not the only taxonomy proposed in the XAI literature [4,6], to the best of the authors' knowledge, it provides the most suitable framework for organizing current research on explainability in point cloud perception for autonomous driving. Nevertheless, like many other taxonomies, it does not correctly distinguish between explainability and interpretability, a distinction that remains conceptually debated.

To clarify this difference, we adopt the definitions provided by Naranjo et al. [23]:

- **Explaining Method**: Most commonly known post-hoc explanations. These methods usually generate explanations to interpret black-box models (Hassija et al. [15]). Explanations can be local (for single input values, e.g. saliency maps) or global (which features contribute more to the overall model)
- **Interpretable Design**: This is the kind of mechanism that comes implicitly in interpretable or white-box models. The functioning of this kind of model is designed to be clear so that the explanation of the output is the model itself. e.g. classification trees.

The primary focus of this survey is on explanation methods rather than on inherently interpretable design. While several studies propose architectures that are inherently interpretable for point cloud perception [16, 32], these works are already addressed comprehensively in other reviews [12, 19, 34]. By focusing on explanation techniques applied to perception models, this survey maintains a clear and targeted scope.

Finally, it is important to acknowledge that XAI remains an evolving field, and certain concepts are still debated regarding whether they should be considered explanations. A notable example concerns attention mechanisms. According to the taxonomy adopted in this survey, and following Wiegreffe et al. [37], attention mechanisms can be regarded as explanatory components. Consequently, they are included within the scope of this work.

3 Explainable Point Cloud Perception in Autonomous Driving

With the recent growth of explainable artificial intelligence (XAI), new approaches have emerged aiming to make 3D perception models more explainable. In this section, we review the state-of-the-art methods and categorize them according to the taxonomy defined in Section 2.

3.1 Hybrid explainable models

The work introduced by Nunes et al. [25] proposes a hybrid pipeline in which a learned black-box representation is used to guide a structured white-box optimization model. The method first extracts point-wise embeddings using a self-supervised deep network (MinkUNet [8] trained with SegContrast [26]). Given an initial instance proposal, the authors compute an attribution map by taking the gradient of the mean feature vector of the proposal with respect to the input features in a local region of interest. Concretely, they average the feature embeddings of the proposal points and backpropagate this scalar quantity to obtain per-point gradients, which quantify how strongly each point contributes to the proposal representation. These gradient magnitudes act as attribution scores: high-saliency points are interpreted as belonging to the foreground instance, while low-saliency points are more likely background. The quality of this attribution map is validated by visaually comparing it to saliency derived from a supervised semantic model and through the downstream segmentation performance.

This attribution map is then used to parameterize a white-box GraphCut formulation. Specifically, high- and low-saliency points are sampled as foreground and background seeds, respectively, defining terminal edge weights in a graph whose non-terminal edges are weighted according to feature-space dissimilarity between neighboring points. The final segmentation is obtained via a minimum cut, yielding a transparent energy-based partition with explicit unary (seed-based) and pairwise (feature-affinity) terms.

Importantly, the authors address the common limitation of white-box models, reduced predictive performance, through benchmarking on the SemanticKITTI dataset [5]. Using the class-agnostic association metric, they demonstrate that their hybrid approach achieves performance competitive with, and in some cases superior to, supervised deep models, particularly on unknown object classes.

3.2 Explainability through architectural adjustments

Several works have incorporated Attention Mechanisms into the training process to enhance both the performance and interpretability of 3D object detectors.

AGS-SSD [29], for instance, improves point cloud downsampling by integrating a Point External Attention (PEA) module. Unlike conventional self-attention, which has a computational complexity of $O(N^2)$, PEA leverages two learnable and shared external memory units (Key and Value) that act as a global feature dictionary. This design enables the generation of a semantic attention map that estimates the importance of each point in the scene. The resulting attention scores are then used to guide the Attention-guided Farthest Point Sampling (A-FPS) algorithm. Instead of relying solely on Euclidean distance, A-FPS adopts a weighted distance metric in which attention scores bias the sampling process toward foreground regions. As a result, semantically relevant object points are preserved during data reduction, mitigating information loss.

SGSSD [33] addresses the information loss and localization inaccuracies introduced by standard downsampling strategies. It proposes a Segmentation-Guided Auxiliary Network (SGAN) that supervises feature learning during training. SGAN jointly performs point-wise semantic segmentation and object center estimation. By predicting whether each point belongs to an object and estimating its offset to the object center, the auxiliary branch encourages the backbone to learn richer structural and semantic representations. This branch is only active during training and is removed at inference time, thereby improving discriminative capability without affecting real-time performance.

To further refine feature representation, SGSSD incorporates a Point Cloud External Attention (PCEA) module. Similar to PEA, PCEA avoids the quadratic complexity of self-attention by employing two lightweight learnable linear layers that serve as external memory units. These units capture global contextual relationships across the entire point cloud rather than restricting interactions to local neighborhoods. The extracted global context is fused with multi-scale backbone features to produce a more robust representation of candidate points.

SP-Det [3] also focuses on identifying and prioritizing informative regions within the voxelized space by introducing an auxiliary Saliency Prediction (SP) task. To mitigate the difficulty of learning from sparse data, the authors implement a Label Diffusion strategy. Instead of supervising the network with sparse, binary ground-truth (where only the exact object voxels are labeled as 1), this strategy spreads the supervision signal to neighboring spatial locations. This creates a soft distribution of labels that provides a denser gradient during training, effectively guiding the model to recognize the saliency or importance of regions surrounding the objects.

The saliency prediction is performed across both the 3D voxel space and the 2D Bird's Eye View (BEV) projection. While the 3D branch preserves volumetric details, the BEV branch provides a computationally efficient global perspective. The results are integrated via the Saliency Fusion (SF) module, which utilizes the saliency maps as a progressive attention mechanism to suppress redundant background features. To ensure localization precision, the architecture incorporates a Hybrid-RoI Pooling module that refines initial proposals by concurrently aggregating features from both the 3D voxel volume and the 2D BEV map.

Previously explained methods were evaluated on the KITTI [13] benchmark and compared against state-of-the-art detectors. Experimental results demonstrate that incorporating attention mechanisms consistently improves 3D object detection performance. Table 1 shows the comparison between each method performance. However, these evaluations focus on detection accuracy and do not directly assess the explainability of the models.

Method	Car (IoU@0.7)	Ped(IoU@0.5)	Cyc(IoU@0.5)
AGS-SSD [29]	81.02	38.53	62.15
SGSSD [33]	81.70	39.35	64.11
SP-Det [3]	83.93	38.91	-

Table 1. Attention-guided neural networks' accuracy on the KITTI [13] test set for 3D object detection.

3.3 Attribution methods

Attribution methods aim to quantify the contribution of each input point to the final prediction of the network, highlighting the most influential regions through saliency maps.

Some approaches generate attribution maps to understand the behavior of 3D object detectors in point clouds.

OccAM [30] generates explanations by empirically estimating point importance through the repeated sub-sampling of the input point cloud and measuring the resulting changes in detection similarity. OccAM employs a density-aware voxel-based sampling strategy that dynamically adjusts the point removal probability to challenge the detector appropriately across varying ranges. Similarly, **3D-SVAM** [18] introduces game-theory by computing a Shapley Value-based Attribution Map. To avoid the exponential complexity of exact Shapley value computation, 3D-SVAM divides the space into voxel units and estimates expected contributions via Monte Carlo sampling. This formulation distinctly enables the quantification of both positive and negative point attributions, facilitating the identification of background or occlusion points that actively hinder correct detection.

In contrast **FFAM** [20] leverages the internal representations of the detection model. FFAM extracts 3D feature maps from the detector's backbone and applies non-negative matrix factorization (NMF) to uncover latent semantic concepts, aggregating them into a global concept activation map. To achieve instance-level interpretability, this global map is refined using the feature gradients de-

rived from an object-specific loss function, yielding high-quality, object-specific saliency maps. Furthermore, FFAM introduces a voxel query-based upsampling strategy to overcome the scale misalignment caused by the network's downsampling, querying sparse neighbors using Manhattan distance to accurately project the activation map back to the original point cloud resolution.

Extending gradient-based interpretation to semantic segmentation tasks, **pGS-CAM** [17] introduces a framework to generate attribution maps for intermediate activation layers. This method computes the gradient influence by aggregating the network's raw output scores (logits) for a target class across a specific subset of points. Specifically, it calculates the gradients of these summed pre-activation scores with respect to the feature maps, effectively capturing the collective impact of the chosen points. Similar to the scale misalignment addressed in FFAM, the resolution of activation maps in encoder-decoder segmentation architectures decreases significantly. To resolve this, pGS-CAM employs a KDTree-based upsampling algorithm to map the coarse localization values back to the original point cloud resolution.

Validating these explanation methods in 3D perception relies on assessing how effectively the generated maps identify critical points. Most approaches rely on point-dropping strategies implemented through Insertion and Deletion metrics [28], which measure changes in performance when important points are removed or added. For instance, OccAM and 3D-SVAM evaluate explanations by progressively removing salient points within a pseudo ground truth bounding box and tracking the degradation in IoU and confidence. Similarly, pGS-CAM validates the robustness of its attribution maps through high-drop and low-drop experiments, measuring the degradation in class-specific and mean Intersection over Union (IoU) when highly salient points are removed. FFAM also employs Insertion and Deletion curves but extends the evaluation with geometric criteria, including Visual Explanation Accuracy (VEA) [27], which computes point-level IoU with ground truth masks.

While the aforementioned methods focus on extracting attributions directly from 3D architectures, a parallel line of research addresses the inherent challenges of LiDAR data, such as extreme sparsity and lack of texture, by incorporating cross-modal guidance or specialized layers to refine the attribution process

To address the noise generated by sparse data, Daniel et al. [11] apply Grad-CAM [31] to object detection in LiDAR point clouds. They transform the point cloud into a 2D Bird's Eye View (BEV) representation, which is then processed by a Single Shot Detector (SSD) to identify objects. However, due to the sparse nature of LiDAR data, many cells in this BEV grid are empty. This sparsity often causes standard convolutional layers to generate noisy activation maps, corrupting the interpretability of the explanations. To address this, the authors propose a Sparsity Invariant Convolutional layer, which ignores empty cells during its operation, thereby producing cleaner activation maps. Finally, the objectivity scores from all cells corresponding to the target class from the network's head are summed. This cumulative sum forms the single value used to initiate the gradient calculation, leading to more focused explanations.

SalLiDAR [9] introduces a cross-modal knowledge transfer framework to incorporate visual saliency into 3D point cloud understanding. The method uses multiple pre-trained 2D saliency models to generate saliency maps from multi-view images. These maps are merged and then projected into 3D space, assigning a continuous saliency value to each point.

Once the 3D saliency distribution is established, it serves as a supervision signal for an auxiliary Saliency Prediction Branch. This branch is trained to regress the projected saliency values directly from the LiDAR features, such as geometry and intensity, using a Kullback-Leibler (KL) divergence loss to align the predicted and transferred distributions.

To validate SalLiDAR the authors integrate the pre-trained saliency network into a two-stream 3D semantic segmentation model and evaluate it on the SemanticKITTI dataset. They explore using the saliency map as an attention-guided loss weighting, as an input feature descriptor, or a combination of both. The results prove that providing the network with this auxiliary saliency information quantitatively improves the mean Intersection-over-Union (mIoU) across various baseline models

Similarly IGO-PQA [41] generates an image-guided point cloud saliency map by fusing LiDAR spatial features with RGB image representations. The process begins by extracting an object-oriented 2D saliency map from surrounding cameras to highlight critical semantic features based on ground truth annotations. Concurrently, the 3D point cloud is evaluated based on its Cartesian distance to the sensor, assigning higher spatial weights to distant points that are inherently harder to capture. By projecting the point cloud onto the image plane, the method calculates the 3D saliency by multiplying each point's normalized distance by its overlapping 2D image saliency intensity. Finally, to account for the inherent sparsity of LiDAR data, a Gaussian pooling process distributes these localized saliency values to neighboring points within a defined radius, constructing a cohesive point-level saliency map.

The aggregated intensities of this point cloud saliency map are then summed and normalized to establish a single, comprehensive quality score for the entire LiDAR frame. This score is then used as a ground truth to train a Neural Network that tries to predict the assigned score. The score itself is not an explanation, however they validate the quality of assigned score by evaluationg 3D object detection models against 3 splits of the nuScenes dataset. The validation dataset is divided into the splits based on the quality score. Results prove that detection models achieve better results in the split with higher score point clouds and worse results in the split with lower score point clouds.

A clear distinction exists in the validation strategies of these two groups. While model-intrinsic methods (e.g., OccAM, FFAM) prioritize quantitative evaluation through perturbation-based metrics like Insertion and Deletion, cross-modal approaches [9, 11, 41] primarily rely on visual analysis or indirect performance gains to justify their saliency maps. This underscores a need for standardized 3D-specific benchmarks that can bridge the gap between qualitative saliency and quantitative faithfulness.

3.4 Example-based explanation methods

Adversarial attacks represent a widely developed field within point cloud perception [14]. While most strategies focus on optimizing a general perturbation to deceive the model, more recent approaches prioritize first identifying why the model makes a particular decision. Consequently, some methods have begun to combine XAI techniques with adversarial attacks to uncover the most relevant features within point clouds.

CAMGA [7] shifts the adversarial focus from the targeted object to its surrounding context in order to preserve the structural integrity of the vehicle. It generates Contextual Attribution Maps by dividing the expanded context area into subregions and measuring the effect of removing each one on the detector's output. The most influential regions are then perturbed using a dual-loss optimization framework that suppresses the detector's output while constraining the perturbation magnitude to remain visually imperceptible.

Similarly, SR-Adv [43] aims to reduce the perturbation budget by identifying critical object regions. It uses Shapley values to assign saliency scores to local regions of the point cloud according to their marginal contribution to the model prediction. The attack is then restricted to the most salient regions, allowing the method to perturb the most vulnerable structures while keeping the modification spatially limited.

GSVA [38] addresses voxel-based 3D detectors, where non-differentiable preprocessing makes direct point-level attacks difficult. Instead of manipulating raw points, the method perturbs voxelized features through re-voxelization or a light voxel attack framework. To preserve sparsity and interpretability, GSVA uses gradient-based attention to identify the most effective voxels, applying adversarial modifications only to highly salient regions.

Wang et al. [35] also rely on point-wise saliency maps to identify vulnerable object regions. Each point is scored using the gradient of the classification loss, and salient points are grouped using Farthest Point Sampling and K-Nearest Neighbors. The resulting local point sets are then perturbed through a dual-loss framework that aims to mislead classification, localization, and confidence estimation while keeping the modified points close to their original positions.

To properly evaluate the effectiveness of adversarial attacks, CAMGA, SR-Adv, and Wang et al. primarily rely on the Attack Success Rate (ASR). ASR measures the proportion of originally correct predictions that the model gets wrong after the perturbation is applied. Alternatively, GSVA measures success using the Relative Corruption Error (RCE) alongside standard detection metrics like Average Precision (AP) or mean AP (mAP). The RCE quantifies the percentage drop in the model's performance by comparing the precision on clean samples against the precision on adversarial samples.

Despite these shared concepts, establishing a direct quantitative comparison between these methods is fundamentally unfeasible. Even though they all utilize the KITTI dataset for their evaluations, the 3D object detection model they use is different. Thus, a direct comparison would be unfair.

3.5 Category overlaps and unrepresented categories

The classification of XAI methods is rarely discrete, as several works combine model-intrinsic and post-hoc strategies. For instance, Nunes et al. [25] is primarily a *Hybrid Explainable Model*, but it uses attribution maps to guide its white-box optimization. SalLiDAR [9] combines attribution, an auxiliary saliency branch, and saliency-oriented regularization. Similarly, adversarial methods such as CAMGA [7], SR-Adv [43], GSVA [38], and Wang et al. [35] are categorized as *Example-based Explanation Methods*, but also rely on attribution mechanisms to identify critical regions or points before applying perturbations. Game-theoretic attribution is also present in 3D-SVAM [18] and SR-Adv [43], which employ Shapley values to estimate feature relevance. Table 2 summarizes the reviewed methods, their main XAI category, additional overlapping categories, and perception task.

An open question is whether combining multiple XAI strategies leads to more explainable or more accurate perception models. Although hybrid approaches may provide complementary insights, current works rarely compare them against single-technique methods under standardized protocols. Therefore, it remains unclear whether methodological overlap translates into measurable gains in explainability or performance.

Reference	Main XAI category	Additional XAI categories	Perception task
[25]	Hybrid explainable models	Attribution Methods	3D Segmentation
[29]	Explainability through architectural adjustments	-	3D Object detection
[33]	Explainability through architectural adjustments	-	3D Object detection
[3]	Explainability through architectural adjustments	-	3D Object detection
[30]	Attribution methods	-	Attribution/Saliency map generation
[18]	Attribution methods	Game Theory	Attribution/Saliency map generation
[20]	Attribution methods	-	Attribution/Saliency map generation
[17]	Attribution methods	-	Attribution/Saliency map generation
[11]	Attribution methods	-	Attribution/Saliency map generation
[9]	Attribution methods	Explainability through Architectural Adjustments, Explainability through Regularization	Attribution/Saliency map generation
[41]	Attribution methods	-	Point cloud quality prediction
[7]	Example-based explanation methods	Attribution methods	Adversarial attack for 3D object detection
[43]	Example-based explanation methods	Attribution methods, Game Theory	Adversarial attack for 3D object detection
[38]	Example-based explanation methods	Attribution methods	Adversarial attack for 3D object detection
[35]	Example-based explanation methods	Attribution methods	Adversarial attack for 3D object detection

Table 2. Summary of existing articles on Explainable Artificial Intelligence (XAI) for Point Cloud Perception in Autonomous Driving.

Despite the rapid growth of the field, several categories from the established XAI taxonomy remain largely unexplored in LiDAR-based autonomous driving:

- *Joint prediction and explanation*
- *Knowledge extraction methods*
- *Visualization methods*
- *Neural methods*

4 Challenges and Limitations

XAI still faces several general challenges that affect its adoption in the broader literature. Previous studies have highlighted the lack of coordination and standardization in both terminology and evaluation protocols, with concepts such as *interpretability* and *explainability* often used interchangeably. This ambiguity is reinforced by the absence of objective and universally accepted metrics

for assessing explanation quality, making it difficult to compare methods fairly. A related issue is explanation fidelity: explanations may appear plausible to humans while failing to accurately reflect the model's actual decision process. Practical deployment also remains challenging, since generating explanations can introduce computational overhead that compromises latency and performance. In offline settings, however, these constraints are less restrictive, making XAI particularly useful for post-hoc reliability analysis, robustness evaluation, and failure diagnosis. Finally, transparency can also introduce vulnerabilities, since information about a model's internal behavior may be exploited to design more effective adversarial attacks, as shown in [7, 35, 38, 43].

In the specific context of point cloud perception for autonomous driving, additional challenges arise from both the maturity of the field and the nature of the data. The intersection of XAI, LiDAR perception, and autonomous driving remains relatively underexplored, and many existing methods are adaptations of image-based approaches rather than techniques designed specifically for sparse, unordered, variable-size 3D data. Current work is also strongly dominated by attribution and saliency maps, whose reliability is often assessed through qualitative visualization or indirect performance gains rather than rigorous 3D-specific validation. This limitation reinforces the ongoing debate on whether saliency maps should be considered faithful explanations or merely indicators of model attention [1, 42]. Moreover, several XAI categories remain largely unexplored in this domain, including joint prediction and explanation, visualization methods, knowledge extraction, and neural concept-based methods. Finally, most XAI advances have historically focused on classification, whereas autonomous driving requires more complex perception tasks such as 3D object detection and semantic segmentation. Adapting explanation methods to these tasks, while preserving faithfulness, scalability, and real-time feasibility, remains a major open challenge.

5 Conclusions and Future Trends

This survey presented a systematic overview of XAI methods for point cloud perception in autonomous driving. Using the taxonomy of Ali et al. [2], we organized the current literature and showed that existing XAI approaches can be adapted to the 3D perception domain. The reviewed works reveal a clear predominance of post-hoc methods, particularly attribution and saliency-based techniques that identify influential points, voxels, or regions involved in detection and segmentation decisions.

The analysis also highlights important limitations. The field still lacks standardized 3D-specific evaluation protocols, making it difficult to compare the faithfulness and usefulness of explanations across methods, datasets, and perception tasks. In addition, several XAI categories remain underexplored in LiDAR-based autonomous driving, while many existing approaches continue to rely on adaptations of image-based methods. Future research should therefore prioritize point cloud–native explanation techniques, stronger validation frameworks, and evaluation metrics tailored to the characteristics of sparse 3D data.

In the coming years, post-hoc methods are likely to remain dominant due to their flexibility and compatibility with existing perception models. Attention mechanisms are also expected to become increasingly common, since they can improve both model performance and spatial interpretability. More broadly, the evolution of XAI for point clouds may follow the same trajectory as point cloud perception itself: early methods adapted image-based techniques through projections such as BEV or spherical maps, while later approaches became increasingly native to 3D data. A similar transition is expected for explainability. Finally, the recent growth of vision-language and 3D language models [10,22,24,36,40] opens a promising direction for generating richer, more human-aligned explanations for autonomous driving perception systems.

Acknowledgment

This work was funded by the Horizon Europe programme of the European Union, under grant agreement 101203230 (project CERTAIN). Funded by the European Union.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9525–9536 (2018)
2. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F.: Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion **99**, 101805 (2023)
3. An, P., Duan, Y., Huang, Y., Ma, J., Chen, Y., Wang, L., Yang, Y., Liu, Q.: Sp-det: Leveraging saliency prediction for voxel-based 3d object detection in sparse point cloud. IEEE Transactions on Multimedia (2024)
4. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion (2020)
5. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In: Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV) (2019)
6. Bereska, L., Gavves, E.: Mechanistic interpretability for ai safety - a review. ArXiv **abs/2404.14082** (2024)
7. Cai, M., Wang, X., Sohel, F., Lei, H.: Contextual attribution maps-guided transferable adversarial attack for 3d object detection. Remote Sensing (2024)
8. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)
9. Ding, G., Imamoglu, N., Caglayan, A., Murakawa, M., Nakamura, R.: Sallidar: Saliency knowledge transfer learning for 3d point cloud understanding. In: 33rd British Machine Vision Conference (BMVC) (2022)

10. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Lowis3d: Language-driven open-world instance-level 3d scene understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
11. Dworak, D., Baranowski, J.: Adaptation of grad-cam method to neural network architecture for lidar pointcloud object detection. Energies (2022)
12. Fernández Llorca, D., Hamon, R., Junklewitz, H., Grosse, K., Kunze, L., Seiniger, P., Swaim, R., Reed, N., Alahi, A., Gómez, E., Sánchez, I., Kriston, A.: Testing autonomous vehicles and ai: perspectives and challenges from cybersecurity, transparency, robustness and fairness. European Transport Research Review (2025)
13. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
14. Guesmi, A., Shafique, M.: Navigating threats: A survey of physical adversarial attacks on lidar perception systems in autonomous vehicles (2024)
15. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A.: Interpreting black-box models: A review on explainable artificial intelligence. Cognitive Computation **16**, 45–74 (2023)
16. Iglesias, A., Aranjuelo, N., Javierre, P., Menendez, A., Arganda-Carreras, I., Nieto, M.: A fully interpretable statistical approach for roadside lidar background subtraction. In: 2025 IEEE International Conference on Vehicular Electronics and Safety (ICVES). pp. 34–41 (2025)
17. Kuriyal, A., Kumar, V.: Towards explainable lidar point cloud semantic segmentation via gradient based target localization (2024)
18. Kuroki, M., Yamasaki, T.: Explaining 3d object detection through shapley value-based attribution map. In: 2024 IEEE International Conference on Image Processing (ICIP) (2024)
19. Kuznietsov, A., Gyevnar, B., Wang, C., Peters, S., Albrecht, S.V.: Explainable ai for safe and trustworthy autonomous driving: A systematic review. IEEE Transactions on Intelligent Transportation Systems (2024)
20. Liu, S., Li, B., Fang, Z., Cui, M., Huang, K.: Ffam: Feature factorization activation map for explanation of 3d detectors (2024)
21. Mulawade, R.N., Garth, C., Wiebel, A.: Explainable artificial intelligence (xai) for methods working on point cloud data: A survey. IEEE Access (2024)
22. Mumuni, F., Mumuni, A.: Explainable artificial intelligence (xai): from inherent explainability to large language models. arXiv preprint arXiv:2501.09967 (2025)
23. Naranjo, R., Aranjuelo, N., Nieto, M., Urbieta, I., et al., J.F.: Design and implementation of a data model for ai trustworthiness assessment in ccam. In: Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART (2025)
24. Natarajan, P., Nambiar, A.: Vale: A multimodal visual and language explanation framework for image classifiers using explainable ai and language models. arXiv preprint arXiv:2408.12808 (2024)
25. Nunes, L., Chen, X., Marcuzzi, R., Osep, A., Leal-Taixé, L., Stachniss, C., Behley, J.: Unsupervised class-agnostic instance segmentation of 3d lidar data for autonomous vehicles. IEEE Robotics and Automation Letters (2022)
26. Nunes, L., Marcuzzi, R., Chen, X., Behley, J., Stachniss, C.: Segcontrast: 3d point cloud feature representation learning through self-supervised segment discrimination. IEEE Robotics and Automation Letters (2022)
27. Oramas M., J., Wang, K., Tuytelaars, T.: Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In: International Conference on Learning Representations (ICLR) (2019)

28. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: British Machine Vision Conference (BMVC) (2018)
29. Qian, H., Wu, P., Sun, B., Su, S.: Ags-ssd: Attention-guided sampling for 3d single-stage detector. Electronics (2022)
30. Schinagl, D., Krispel, G., Possegger, H., Roth, P.M., Bischof, H.: Occam's laser: Occlusion-based attribution maps for 3d object detectors on lidar data. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
31. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
32. Wang, D.Z., Posner, I.: Voting for voting in online point cloud object detection. In: Robotics: Science and Systems (2015)
33. Wang, X., Zhang, D., Niu, H., Liu, X.: Segmentation can aid detection: Segmentation-guided single stage detection for 3d point cloud. Electronics (2023)
34. Wang, Y., Mao, Q., Zhu, H., Deng, J., Zhang, Y., Ji, J., Li, H., Zhang, Y.: Multi-modal 3d object detection in autonomous driving: A survey. International Journal of Computer Vision (2023)
35. Wang, Y., Wu, L., Jin, J., Wang, E., Zhang, Z., Zhao, Y.: An imperceptible adversarial attack against 3-d object detectors in autonomous driving. IEEE Internet of Things Journal (2025)
36. Wang, Y., Wang, S., Zhang, Z., Lu, X., Cai, C., Li, H., Liu, F., Jia, P., Lang, X.: Uniplv: Towards label-efficient open-world 3d scene understanding by regional visual language supervision. ArXiv (2024)
37. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 11–20 (2019)
38. Wu, J., Yao, W., Jia, S., Jiang, T., Zhou, W., Ma, C., Chen, X.: Gradient-based sparse voxel attacks on point cloud object detection. Pattern Recognition (2025)
39. Yeong, D.J., Panduru, K., Walsh, J.: Exploring the unseen: A survey of multi-sensor fusion and the role of explainable ai (xai) in autonomous vehicles. Sensors (2025)
40. Zeng, Y., Jiang, C., Mao, J., Han, J., Ye, C., Huang, Q., Yeung, D.Y., Yang, Z., Liang, X., Xu, H.: Clip2: Contrastive language-image-point pretraining from real-world point cloud data. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
41. Zhang, C., Eskandarian, A.: Image-guided outdoor lidar perception quality assessment for autonomous driving. IEEE Transactions on Intelligent Vehicles (2025)
42. Zhang, H., Figueroa, F.T., Hermanns, H.: Saliency Maps Give a False Sense of Explanability to Image Classifiers: An empirical evaluation across methods and metrics. In: Proceedings of the 16th Asian Conference on Machine Learning. vol. 260, pp. 479–494 (2025)
43. Zheng, S., Liu, W., Guo, Y., Zang, Y., Shen, S., Wen, C., Cheng, M., Zhong, P., Wang, C.: Sr-adv: Salient region adversarial attacks on 3d point clouds for autonomous driving. IEEE Transactions on Intelligent Transportation Systems (2024)