NBA3D: Neighbor-Based Confidence Adjustment for 3D Rare Object Detection Using LiDAR

Jooyoung Lee[1], **Jaeyoon Lee**[1], **Jongwon Choi**[1,2*]

[1]Dept. of Artificial Intelligence, Chung-Ang University, Seoul, Korea
[2]Dept. of Advanced Imaging, GSAIM, Chung-Ang University, Seoul, Korea
{juyeoe, leejaeyoon}@vilab.cau.ac.kr, choijw@cau.ac.kr

Abstract

Recent research on LiDAR-based 3D object detectors has shown strong performance; however, evaluations typically focus on dominant classes, overlooking rare classes, such as strollers, which could be critical in real autonomous driving scenarios. This oversight is problematic because state-of-the-art 3D object detectors show significantly lower performance on rare classes compared to dominant ones when trained on both. To address this issue and achieve accurate 3D rare object detection using only LiDAR data, we propose the Neighbor-Based confidence Adjustment for 3D rare class predictions (NBA3D). NBA3D utilizes a graph neural network to analyze the surrounding environment of rare class prediction boxes, enabling a more effective distinction between true positives and false positives based on their local context. Our approach utilizes both 3D prediction box characteristics and CLIP-based class semantic information to better contextualize neighboring objects. Various experiments demonstrate that NBA3D effectively improves the detection performance of rare class objects, regardless of the type of 3D object detectors used.

Introduction

Among the various sensors used for 3D object detection in autonomous vehicles, LiDAR stands out due to its robustness against varying weather and lighting conditions, precise distance measurement, and accurate object shape identification. Autonomous vehicles rely on data from multiple sensors, including cameras, Radar, and LiDAR, to navigate safely through diverse and unpredictable road environments (Mao et al. 2023). While cameras offer rich visual information, they are susceptible to weather and lighting challenges and struggle with accurate distance estimation. Radar, which uses electromagnetic waves, provides reliable detection across various conditions but is limited in object identification due to its low resolution. In contrast, LiDAR excels in both distance measurement and detailed object shape recognition by generating 3D point clouds through laser pulses, making it a crucial sensor in autonomous driving systems.

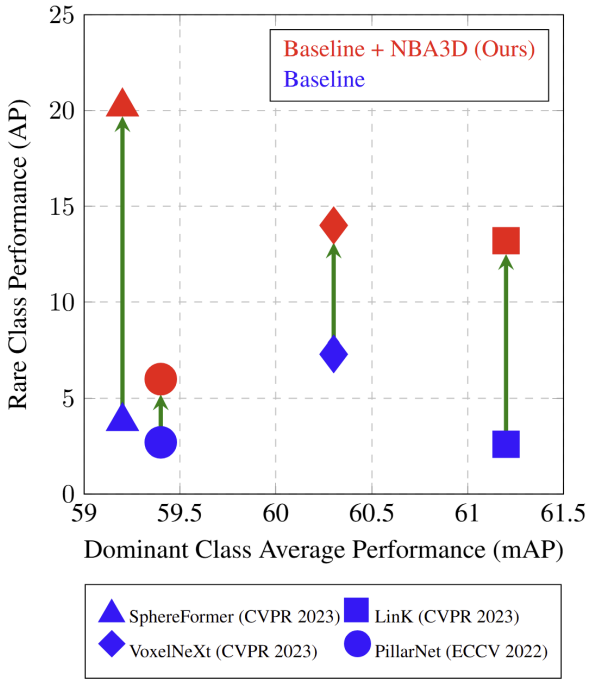

Figure 1: Performance Before and After NBA3D Application. Regardless of the type of state-of-the-art LiDAR-based 3D object detector, our NBA3D significantly improves the performance of the Stroller class, which is one of the rare classes while maintaining performance on dominant classes.

Although state-of-the-art LiDAR-based 3D object detectors (Zhan et al. 2023; Chen et al. 2023c; Zhang et al. 2023) have demonstrated impressive performance, they encounter significant challenges when dealing with a diverse range of objects that follow a long-tail distribution (Liu et al. 2019). This distribution is characterized by a few frequently occurring dominant object categories ('head') and a wide variety of rare objects ('tail'). Despite the prevalence of this distribution in real-world scenarios, current detectors (Lu et al. 2023; Lai et al. 2023; Chen et al. 2023b) commonly focus on evaluating the performance on dominant classes. This focus reveals a critical limitation: when rare categories are trained alongside dominant ones, detection performance on the rare

*Corresponding author

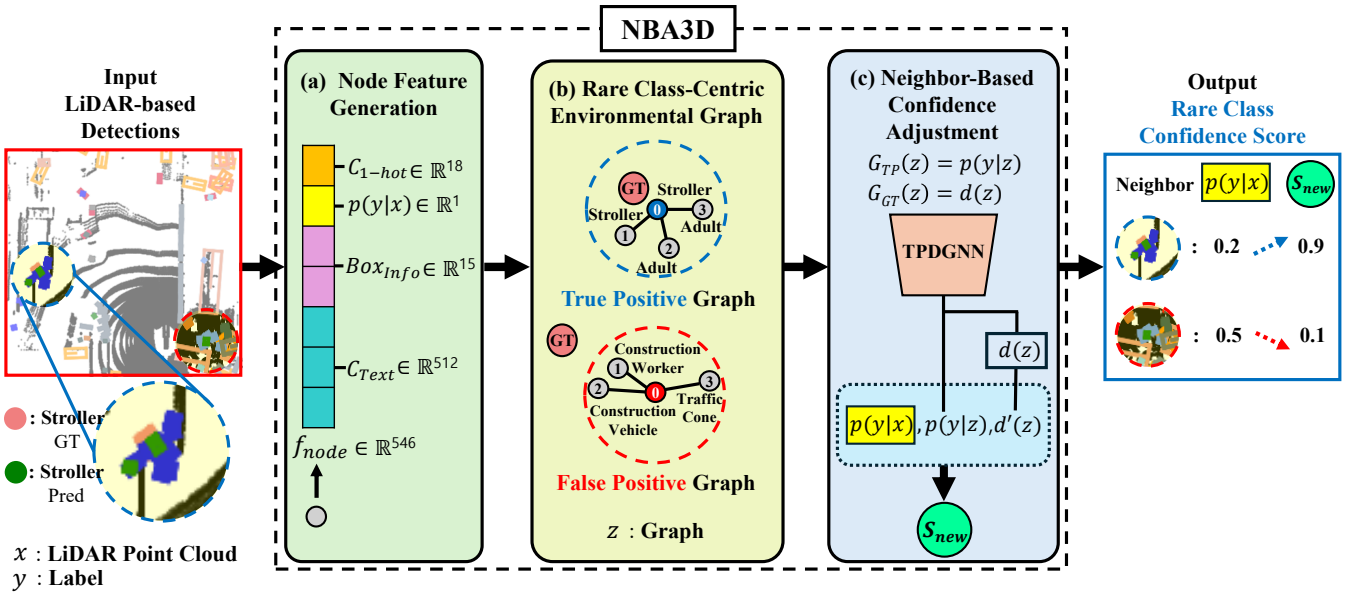

Figure 2: Visualization of the Proposed NBA3D. (a) Prediction boxes obtained from LiDAR-based 3D object detectors are represented as node features consisting of class information C, confidence score $p(y|x)$, and box details. (b) An environmental graph centered on rare class prediction boxes is generated. (c) Through using neighboring information, the confidence score $p(y|x)$ of the central rare class prediction box is reassigned by an adjusted confidence score S_{new}.

objects significantly degrades, posing a serious concern for the safety of autonomous vehicles.

Previous research aimed at improving rare object detection in LiDAR-based 3D object detectors (Peri et al. 2023; Ma et al. 2023) has faced challenges due to the reliance on visual sensors. The previous methods enhanced performance on rare classes by implementing a group-free head, utilizing class hierarchy, and applying filtering based on camera-based prediction boxes. However, these approaches require both camera and LiDAR sensors, which increases system complexity and cost. They also need changes to the 3D object detector model structure to handle multiple modalities, making it difficult to integrate with existing models without substantial modifications. The multi-sensor approach also diminishes the advantages of a LiDAR-only system, such as visual robustness and precise distance measurement. These constraints limit the applicability of the approaches, particularly in scenarios where minimizing hardware requirements or preserving the existing model architecture is critical.

To address these challenges, we propose NBA3D, which is a novel method reassigning confidence scores to rare class prediction boxes only based on their surrounding environment. NBA3D utilizes common-sense knowledge about object relationships, such as the higher likelihood of adults or children being near a stroller rather than trucks or construction vehicles. For this purpose, our approach constructs a graph centered on the prediction boxes, where each node encodes the semantic context of object categories using a CLIP-based text encoder. Crucially, NBA3D relies solely on the information predicted by LiDAR-based 3D object detectors, making it more versatile and easier to integrate with existing models. This approach improves performance while maintaining the baseline detector model without requiring its modifications. Additionally, we introduce a new integration module that combines the original and refined predictions based on the reliability of the constructed graphs. Experimental results on the autonomous vehicle environment demonstrate that NBA3D significantly enhances the performance of LiDAR-based 3D object detectors on rare classes while preserving the performance of dominant ones.

The main contributions of this paper are as follows:

- We propose NBA3D that effectively reassigns confidence scores for rare classes based on the surrounding environment obtained through LiDAR-based 3D object detectors without requiring additional sensors.
- We introduce a graph around each rare class prediction box to refine its confidence score by integrating information from neighboring prediction boxes and class semantic context.
- We design an integration module that combines the original and refined confidence scores, based on the reliability of the given graph.
- Our extensive experiments in vehicle environments show that NBA3D improves the performance of rare classes regardless of the type of 3D object detector used.

Related Work

LiDAR-based 3D Object Detection LiDAR-based 3D Object Detectors typically consist of a backbone, neck, and head structure. Numerous studies (Yan, Mao, and Li 2018) have focused on improving each component to enhance performance while reducing computational costs. Unlike 2D images composed of pixel grids, point clouds obtained

from LiDAR consist of sets of points with 3D coordinates where the order of points is not important. Therefore, a preprocessing procedure is necessary before the backbone can process the point cloud input using CNNs. Although point-based methods (Qi et al. 2017; Shi, Wang, and Li 2019) using farthest point sampling and graph-based approaches (Shi and Rajkumar 2020) exist, voxelization (Zhou and Tuzel 2018) and pillar-based approaches (Lang et al. 2019) have proven their efficiency for processing LiDAR point cloud data in 3D object detection tasks by representing the data as 3D voxels or 2D column-like structures, respectively.

CenterPoint (Yin, Zhou, and Krahenbuhl 2021) used an anchor-free approach with a CenterHead, predicting heatmaps where object centers have a value of 1. Building upon the idea of efficiently utilizing large kernels (Chen et al. 2023a), LinK (Lu et al. 2023) introduced a linear kernel generator to effectively expand the receptive field. PillarNet (Shi, Li, and Ma 2022) utilized a feature fusion neck. Considering the sparse nature of points in distant regions and dense points in nearby areas (Qian, Lai, and Li 2022), SphereFormer (Lai et al. 2023) leverages radial window self-attention. GeoMAE (Tian et al. 2023) fine-tunes a transformer-based point cloud encoder (Fan et al. 2022) pretrained on centroid prediction, normal estimation, curvature prediction, and occupancy prediction.

However, traditional methods tend to focus on dominant classes, leading to significantly lower performance for rare classes like 'Stroller' compared to more common classes like 'Car.' Our approach enhances state-of-the-art 3D object detectors to account for rare categories within a long-tail distribution, thereby improving detection accuracy for these rare objects without compromising performance on dominant classes.

Rare Object 3D Detection Peri *et al.* (Peri et al. 2023) proposed a solution for LiDAR-based 3D object detectors when training on both dominant and rare classes. Following CBGS (Zhu et al. 2019), these detectors typically use a multi-group head that groups classes with similar shapes. However, the process of non-maximal suppression (NMS) can lead to the deletion of rare class prediction boxes with low confidence scores due to dominant class prediction boxes in the same group. To prevent the issue, Peri *et al.* (Peri et al. 2023) suggested a group-free head where NMS is applied only within the same class.

Although Peri *et al.* (Peri et al. 2023) improved performance on rare classes, they required both camera and LiDAR data and necessitate changes to the head structure of the 3D object detector model to integrate multiple modalities. In contrast, our NBA3D enhances rare class performance without modifying the architecture of existing LiDAR-based 3D object detector models or requiring additional camera data. NBA3D is applied to the prediction box results of existing LiDAR-based 3D object detectors, offering a versatile and efficient solution to improve rare class detection.

Method

In this section, we begin with an overview of the NBA3D architecture. Subsequently, we provide an in-depth explanation for the modules of node feature generation, rare class-centric environmental graph, true positive discrimination graph neural network, and neighbor-based confidence adjustment.

Preliminary

A LiDAR-based 3D object detector takes LiDAR point cloud frames as an input x and outputs predicted bounding boxes for each class. For each bounding box, the model predicts the one-hot encoded class $C_{\text{1-hot}}$, the confidence score $p(y|x)$ for label y, and the box information Box_{Info} which includes center coordinates, dimensions, rotation, relative center coordinates to the ego vehicle, and velocity. The Box_{Info} derived from training data is used to train NBA3D. Then, NBA3D reassigns confidence scores for rare class prediction boxes on a frame-by-frame basis using the Box_{Info} from the test data.

Proposed Method

The architectural framework of NBA3D is illustrated in Fig. 2. The process begins by obtaining prediction boxes for the training set using a LiDAR-based 3D object baseline detector. Our framework focuses specifically on environmental graphs centered around predictions of rare classes, which are specified from the training set. During inference, NBA3D selectively processes only these rare class predictions.

For each frame, we generate an environmental graph centered around each rare class prediction box. We then create node features for each graph and use a Graph Neural Network (GNN) to compute new confidence scores, which are reassigned to the corresponding rare class prediction boxes. We call the GNN by *True Positive Discrimination GNN* (TPDGNN).

Node Feature Generation As shown in Fig. 2 (a), we represent each predicted bounding box with a node feature vector f_{node}. The node features f_{node} consist of four key elements: an 18-dimensional one-hot class vector $C_{\text{1-hot}}$, a single confidence score, box information Box_{Info}, and the class text embedding C_{Text}. The confidence score $p(y_i|x)$ reflects the likelihood of the label y_i given the LiDAR point cloud x. The Box_{Info} contains 15 values: 3 for the x, y, z center coordinates, 3 for the object's length, width, and height, 4 for the heading angle rotation quaternion, 3 for the center coordinates relative to the ego vehicle, and 2 for the velocity in the x and y directions. This aggregated feature set captures essential information about object classification, detection confidence, spatial properties, and kinematic characteristics in 3D space. The class text embedding C_{Text} is obtained from the category name through a CLIP-based text encoder. For instance, the text embedding for a truck is obtained by applying L2-normalization to the encoded vector from a pre-trained CLIP (Radford et al. 2021) text encoder (ViT-B/32 version) fed by 'vehicle.truck.'

Thus, we can denote f_{node} as follows:

$$f_{node} = [C_{\text{1-hot}}, p(y_i|x), Box_{\text{Info}}, C_{\text{Text}}]\,. \tag{1}$$

Rare Class-Centric Environmental Graph As illustrated in Fig. 2 (b), for each frame, we construct an environmental graph centered on each predicted bounding box of rare class objects. The central rare class prediction box becomes the 0 node, and all prediction boxes within a 4-meter radius of its center coordinate become neighboring nodes. The distance between the center coordinates of each neighboring box and the central prediction box is used as the edge feature f_{edge}. If no prediction boxes exist within the 4-meter radius, the five nearest prediction boxes to the central rare class prediction box are selected as neighbors.

The label of the graph is determined as follows: If the distance between the center of the target prediction box and the nearest rare class ground truth box is less than 4 meters, the target box is labeled as a true positive, setting the graph's label to 1. If the distance is larger than 4 meters or if no rare class ground truth box exists in the frame, the target box is labeled as a false positive, and the graph's label is set to 0. The input to the True Positive Discrimination GNN (TPDGNN) consists of the environmental graph centered on the rare class, the graph's label, and the distance between the center of the target prediction box and its nearest rare class ground truth box. If there is no rare class ground truth box in the frame, the distance between the prediction and ground truth boxes is set to 1000 meters as an infinity.

To achieve effective normalization across both the training and test data, we first compute the normalization parameters exclusively from the training data. For node features f_{node}, we normalize the confidence score $p(y|x)$ and Box_{Info} components excluding rotation quaternion by subtracting the mean and dividing by the standard deviation of their respective distributions computed from the training data. Specifically, we calculate minimum (min) and maximum (max) values of edge features f_{edge} from the training data. Additionally, we determine the minimum and maximum distances between the center of the rare class prediction box and their nearest rare class ground truth box, excluding any outlier distances that were artificially set to 1000 meters (i.e., infinity). These computed values are then used to constrain all distances within a range defined by the calculated minimum d_{min} and maximum d_{max} values. For the final normalization step d_{norm}, distances less than or equal to 4 meters are linearly normalized into the range of 0 to 0.5, while distances greater than 4 meters are also linearly normalized into the range of 0.5 to 1 where the maximum distance within the entire training data is normalized.

We can derive the normalization of each distance d_i as follows:

$$\begin{aligned} d_{norm} =& 0.5 \cdot min\left(1, max\left(0, \frac{d_i - d_{min}}{4 - d_{min}}\right)\right) \\ &+ 0.5 \cdot max\left(0, min\left(1, \frac{d_i - 4}{d_{max} - 4}\right)\right) \end{aligned} \tag{2}$$

True Positive Discrimination GNN Fig. 3 illustrates the structure of the TPDGNN. The network receives input with a batch size of B, N nodes, and E edges: node features f_{node} of size $B \times N \times 546$, edge indices E_{idx} of size $B \times 2 \times E$, and edge features f_{edge} of size $B \times E \times 1$.

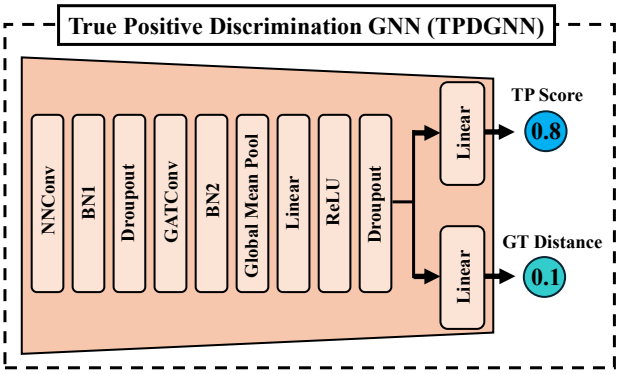

Figure 3: Composition of True Positive Discrimination GNN (TPDGNN).

The process begins with the $NNConv$ layer (Gilmer et al. 2017), which transforms the edge features f_{edge} to a size of $B \times E \times 8736$. This 8,736 is the result of multiplying the node feature f_{node} dimension 546 by 16. This transformation enriches the edge information, enhancing the representation of relationships between nodes. With this enriched edge information, the node features f_{node} are updated to $B \times N \times 16$, aggregating neighboring node information by taking the mean, thereby capturing local structures. Next, batch normalization BN is applied to the node features of size $\times N \times 16$ to stabilize the learning process, followed by a 30% dropout D to prevent overfitting, resulting in the output H_1 as follows:

$$H_1 = D(BN(NNConv(f_{node}, E_{idx}, f_{edge}))). \tag{3}$$

Subsequently, the Graph Attention (GAT) layer (Veličković et al. 2017) takes the $B \times N \times 16$ input H_1 and generates a $B \times N \times 16$ output. During this process, GAT layer learns the importance between nodes, assigning higher weights to more significant neighbors. Batch normalization BN and dropout D ($p = 0.3$) are applied again to further enhance learning stability. Global mean pooling GMP transforms the node-level features of size $B \times N \times 16$ into a $B \times 16$ graph-level representation H_2 of the environment surrounding the central rare class as follows:

$$H_2 = GMP(D(BN(GAT(H_1)))). \tag{4}$$

This formulation has the effect of summarizing the overall characteristics of each surrounding environment graph.

Then, through a fully connected layer, the H_2 representation is expanded from $B \times 16$ to $B \times 256$, which increases the model's expressive power, allowing it to learn more complex patterns. ReLU activation and dropout ($p = 0.3$) are also applied at this stage. Finally, we use two separate layers fed by the expanded feature, which are G_{TP} and G_{GT} respectively for distinct predictions. When z represents the environmental graph surrounding the center of a predicted box for rare classes, G_{TP} generates a true positive score prediction $p(y_i|z_i)$ of size $B \times 1$, while G_{GT} outputs a ground truth distance prediction $d(z_i)$ of the same size. Thus,

$$G_{TP} = p(y_i|z_i), G_{GT} = d(z_i). \tag{5}$$

Method	Many	Medium	Rare	Pedestrian	Debris	Stroller	Emergency Vehicle	Personal Mobility
SphereFormer	74.7	46.3	5.7	83.6	5.0	3.8	13.5	0.4
+ NBA3D (Ours)	74.7	46.3	10.8	83.9	6.6	20.2	14.1	2.4
+ Improvement	+0.0	+0.0	+5.1	+0.3	+1.6	+16.4	+0.6	+2.0
LinK	75.5	49.2	6.4	84.3	5.6	2.6	16.9	0.6
+ NBA3D (Ours)	75.6	49.2	9.4	84.5	6.2	13.2	17.1	1.0
+ Improvement	+0.1	+0.0	+3.0	+0.2	+0.6	+10.6	+0.2	+0.4
VoxelNeXt	74.4	48.5	8.2	84.9	5.6	7.3	16.1	3.6
+ NBA3D (Ours)	74.4	48.5	12.0	85.0	6.5	14.0	16.2	11.3
+ Improvement	+0.0	+0.0	+3.8	+0.1	+0.9	+6.7	+0.1	+7.7
PillarNet	75.3	46.1	6.4	83.2	4.8	2.7	17.7	0.2
+ NBA3D (Ours)	75.3	46.1	7.6	83.2	5.1	6.0	18.1	1.1
+ Improvement	+0.0	+0.0	+1.2	+0.0	+0.3	+3.3	+0.4	+0.9

Table 1: Impact on Rare Class Performance of State-of-the-Art LiDAR 3D Detector After NBA3D Application. 'Many' represents the mAP of the top 5 classes by training instance count: Car, Pedestrian, Barrier, Traffic Cone, and Truck. 'Medium' represents the mAP of the next 6 classes: Trailer, Pushable Pullable, Bus, Construction Vehicle, Motorcycle, and Bicycle. Adult, Construction Worker, Child, and Police Officer are classified under the same Pedestrian class.

This dual-layer structure enables the model to perform classification and regression tasks simultaneously, allowing the model to learn about two related tasks by utilizing the structural information of the graph. The complex structural information of the surrounding environment graphs is progressively compressed and transformed to effectively distinguish between true positives and false positives.

Loss Function The TPDGNN is trained using a loss function L that combines two components: a sigmoid focal loss term L_{TP} (Lin et al. 2017) and a smooth L1 loss term L_{GT} (Girshick 2015) as follows:

$$L = L_{TP}(p(y|z), s) + \lambda_{GT} L_{GT}(d(z), d). \quad (6)$$

L_{TP} distinguishes whether the central rare class prediction box is a true positive or a false positive by analyzing the surrounding environment of rare class prediction boxes. L_{GT} predicts the distance between the center of the central rare class prediction box and the center of the nearest rare class ground truth box. The balance between these two loss terms is controlled by a hyperparameter λ_{GT}, which is set to 0.5. In the sigmoid focal loss, the weight alpha for the true positive label (1) is set to 0.8. The gamma parameter, which gives higher weight to difficult examples that are misclassified, thereby focusing on challenging cases, is set to 3.

Neighbor-Based Confidence Adjustment As shown in Fig. 2 (c), the new confidence score S_{new} for the central rare class prediction box is computed as the average of three values: the existing confidence score $p(y_i|x)$, the true positive score $p(y_i|z_i)$ obtained through the TPDGNN, and a confidence score based on the predicted distance to the center coordinates of the nearest rare class ground truth box $d'(z_i)$. Thus, we can update the original confidence score by S_{new} estimated by:

$$S_{\text{new}} = \text{mean}(p(y_i|x), p(y_i|z_i), d'(z_i)), \quad (7)$$

where $d'(z_i) = \max(0, -2[d(z_i)]_0^1 + 1)$ and $[\bullet]_0^1$ represents a clipping function from 0 to 1.

Experiments

To demonstrate the effectiveness of NBA3D, we reassign the confidence scores of rare class prediction boxes from various state-of-the-art LiDAR-based 3D object detectors using neighbor-based confidence, and evaluate the results. We apply this method to models that perform well on 10 dominant classes, including SphereFormer (Lai et al. 2023), PillarNet (Shi, Li, and Ma 2022), VoxelNext (Chen et al. 2023b), and LinK (Lu et al. 2023).

Experimental Setup

Dataset. To better reflect real-world autonomous driving scenarios with imbalanced class distributions, we employ the long-tail variant of the nuScenes dataset (Caesar et al. 2020). Therefore, following the approach of Peri *et al.* (Peri et al. 2023), we utilize a total of 18 long-tail classes to provide a more comprehensive evaluation. The rare classes are defined as those with fewer than 5k training instances per class.

Evaluation Metric. While extending the original nuScenes dataset to 18 classes, we utilize the same rigorous evaluation protocols which provide a more challenging and realistic testbed for rare object detection. For each class, the average precision (AP) is the average of APs computed at center distance thresholds of 0.5, 1, 2, 4 meters. Consistent with the methodology of Peri *et al.* (Peri et al. 2023), our training employs the nuScenes train set, while evaluation is conducted on the val set, covering all 18 classes.

Implementation Details. In our experiments, we use published code for all LiDAR-based 3D object detector base models, training them on 18 classes of data. While adhering to default configurations and training policies, we adjust the learning rate proportional to the batch size as suggested by Goyal et al. (Goyal et al. 2017). Specifically, LinK (Lu et al. 2023) is trained with two 3090 GPUs, using a total batch size of 8. SphereFormer (Lai et al. 2023) utilizes four A6000 GPUs with a total batch size of 8. VoxelNeXt (Chen

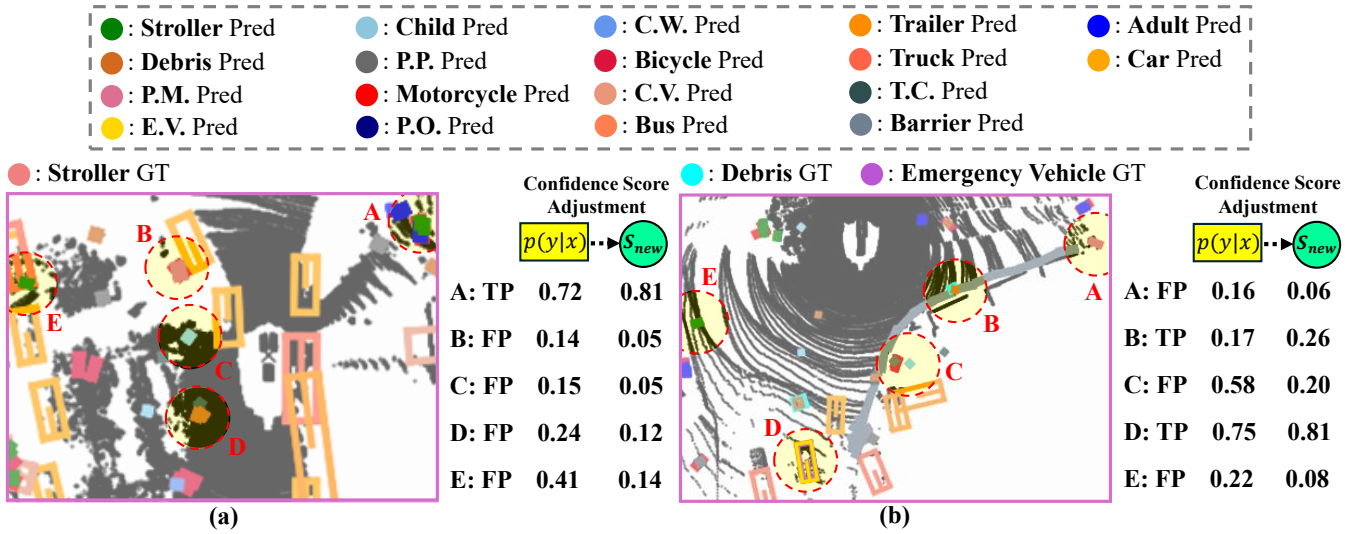

Figure 4: Adjusted Rare Class Confidence Score Based on NBA3D. Prediction boxes for each class and rare class ground truth boxes are distinguished by color. NBA3D analyzes the surrounding environment of the rare class prediction box center and reassigns the confidence score of the central rare class prediction box. $p(y|x)$ represents the original confidence score of the central rare class prediction box, while S_{new} is the adjusted confidence score after applying NBA3D.

Base	One-Hot, Conf., Box Information	Text Embedding	Pedestrian	Debris	Stroller	E.V.	P.M.	Rare
✓	-	-	83.6	5.0	3.8	13.5	0.4	5.7
✓	✓	-	84.1	6.1	6.1	14.4	1.2	7.0
✓	✓	✓	83.9	6.6	20.2	14.1	2.4	10.8

Table 2: NBA3D Performance Ablation on Node Feature Configurations. Using SphereFormer as the base detector, we demonstrate the performance changes of NBA3D resulting from the addition of node feature configuration information. 'Conf.' stands for the original confidence score of prediction boxes.

et al. 2023b) employs a single Titan GPU with a batch size of 4. PillarNet (Shi, Li, and Ma 2022) uses one 3090 GPU with a batch size of 8. Regardless of the LiDAR-based 3D object detector type, NBA3D consistently uses a single A6000 GPU with a batch size of 256 for 10 epochs. Due to the higher prevalence of false positive graphs with label 0, we replicate the true positive graphs with label 1 to match the number of label 0 graphs. In each batch of 256, we maintain an equal distribution. We set the learning rate to 0.04 and use the Adam optimizer (Kingma and Ba 2014).

Through optimized graph construction and streamlined architecture, NBA3D introduces minimal computational overhead (0.31ms for graph generation, 0.61ms for feature normalization, 0.05ms for GNN inference), which enhances its usability with existing 3D detectors.[1]

Experimental Results

Quantitative Comparisons. Table 1 demonstrates that the application of NBA3D consistently improves the performance on rare classes across various LiDAR-based 3D object detector models. Notably, the performance enhancement for the stroller class is particularly impressive. NBA3D only reassigns confidence scores for prediction boxes of rare classes, thus maintaining performance on the 11 dominant classes without degradation.

Change in Rare Class Confidence Score. Fig. 4 shows the confidence scores of rare class prediction boxes before and after applying NBA3D. Using NBA3D, which utilizes the surrounding environment of rare class prediction boxes, the confidence scores of true positive prediction boxes close to the ground truth rare class boxes within the frame increase to approach 1, while the confidence scores of distant false positive prediction boxes and rare class prediction boxes in frames without ground truth rare class boxes decrease to approach 0. This approach enhances the reliability of certainty predictions.

Ablation Study

To evaluate the effectiveness of the proposed node feature generation, rare class-centric environmental graph, and neighbor-based confidence adjustment, we assess the impact of each variation on NBA3D's performance using the nuScenes validation set.

Effects of Node Feature Information Table 2 demonstrates that incorporating class text embeddings into the node features representing each prediction box contributes to NBA3D's performance improvement, as it takes into

[1]Project page: https://juuyeoe.github.io/NBA3D/

B	Max	Ped.	Deb.	Str.	E.V.	P.M.	Rare
✓	-	83.6	5.0	3.8	13.5	0.4	5.7
✓	5	83.9	5.6	7.8	14.3	0.6	7.1
✓	10	83.7	6.3	13.6	13.6	2.9	9.1
✓	All	83.9	6.6	20.2	14.1	2.4	10.8

Table 3: NBA3D Performance Ablation on Maximum Neighbor Count. 'B' denotes the baseline detector performance, while 'All' is a variant where all prediction boxes within the radius are considered as neighbors.

L_{TP}	L_{GT}	Ped.	Deb.	Str.	E.V.	P.M.	Rare
-	-	83.6	5.0	3.8	13.5	0.4	5.7
✓	-	83.4	6.7	19.7	13.7	0.6	10.2
✓	✓	83.9	6.6	20.2	14.1	2.4	10.8

Table 4: NBA3D Performance Gains by Loss Configuration Ablation.

account the similarity between class text embeddings. The text embedding for 'human.pedestrian.stroller' utilizes information about the similarities between adult, child, police_officer, construction_worker, and personal_mobility, which all belong to the human.pedestrian category in the nuScenes dataset. It also leverages the vast relationships between text and images learned by the pre-trained CLIP (Radford et al. 2021) text encoder to utilize information from classes similar to stroller. The text embedding for 'human.pedestrian.personal_mobility' implies the context that a personal_mobility is a pedestrian-related object used by humans, containing more information than a one-hot vector and thus contributing to performance improvement.

Effects of Maximum Number of Neighbors Table 3 examines the performance changes as we limit the number of neighbors within a given radius. When all prediction boxes within a 4-meter radius become neighbors of a rare class prediction box at the center, the number of nodes in each graph around the central rare class prediction box varies. If we allow only up to a maximum number of neighbors, prioritizing those closest to the central rare class prediction box, performance is better with a maximum of 10 neighbors compared to 5. However, performance is best when there is no maximum limit, allowing all prediction boxes within the radius to be neighbors, as this utilizes information from diverse environments. Notably, for isolated debris predictions (>4m from neighbors), NBA3D's strategy of using the 5 nearest boxes effectively reduces false positive confidence scores from 0.174 to 0.112, with maximum reduction reaching 0.452.

Effect of Predicting Distance from GT Table 4 compares the effectiveness of using only a loss that learns the graph labels of the environment surrounding prediction boxes centered on rare classes, versus using a combined loss that learns both graph labels and the distance to the nearest ground truth. Simultaneously optimizing classification accuracy and distance prediction promotes complementary learning through efficient extraction of shared features, leading

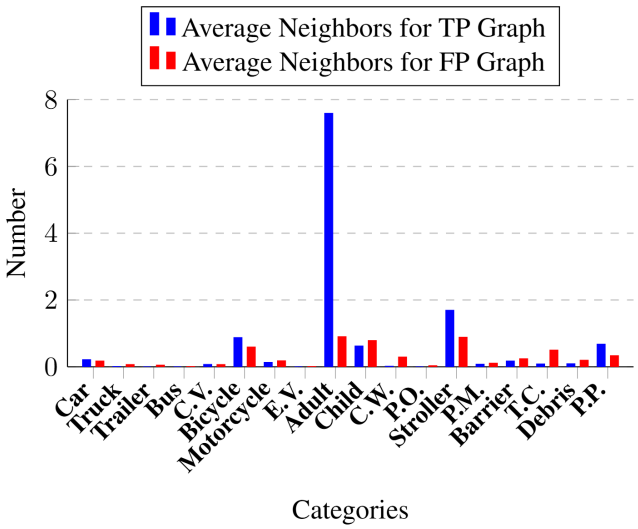

Figure 5: Stroller Vicinity Neighbor Distribution Shifts Across Surrounding Graph Labels.

to more generalized representations that reduce overfitting. This results in more accurate and robust predictions, enhancing NBA3D performance. When not predicting distance from GT, a new confidence score is assigned as the average of the true positive score and the original confidence score.

Class-wise Performance Improvement Analysis

NBA3D provides the most significant improvement in stroller detection performance. Fig. 5 shows that environments where strollers appear exhibit distinct dominant characteristics. In the vicinity of true positive stroller prediction boxes, adults frequently occur. The surroundings of accurately predicted stroller locations consistently show a high presence of adults. Emergency vehicles show moderate gains due to their context diversity.

Conclusion

This paper proposes NBA3D, a method to improve the performance of LiDAR-based 3D object detectors on rare classes. NBA3D generates an environmental graph centered on rare class prediction boxes, connecting them with neighboring predictions within a specified radius. Node features representing prediction boxes incorporate 3D box characteristics and class text embeddings. By leveraging neighboring information, NBA3D distinguishes whether the central rare class prediction box is a true positive or false positive, and reassigns an adjusted confidence score to the central prediction. Extensive experiments demonstrate that NBA3D is effective in enhancing the performance on rare classes. Through this paper, we confirmed that it is possible to overcome the existing data scarcity issue purely by utilizing predicted results, correlations among objects, and their prior information, without employing additional sensors or modifying the architecture.

Acknowledgments

This work was partly supported by National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT (2021M3H2A1038042) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (Ministry of Science and ICT) (2021-0-01341, Artificial Intelligence Graduate School Program (Chung-Ang University); ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00437102))

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; and Jia, J. 2023a. Largekernel3d: Scaling up kernels in 3d sparse cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13488–13498.

Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; and Jia, J. 2023b. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21674–21683.

Chen, Y.; Yu, Z.; Chen, Y.; Lan, S.; Anandkumar, A.; Jia, J.; and Alvarez, J. M. 2023c. Focalformer3d: focusing on hard instance for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8394–8405.

Fan, L.; Pang, Z.; Zhang, T.; Wang, Y.-X.; Zhao, H.; Wang, F.; Wang, N.; and Zhang, Z. 2022. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8458–8468.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, 1263–1272. PMLR.

Girshick, R. 2015. Fast r-cnn. *arXiv preprint arXiv:1504.08083*.

Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lai, X.; Chen, Y.; Lu, F.; Liu, J.; and Jia, J. 2023. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17545–17555.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2537–2546.

Lu, T.; Ding, X.; Liu, H.; Wu, G.; and Wang, L. 2023. Link: Linear kernel for lidar-based 3d perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1105–1115.

Ma, Y.; Peri, N.; Wei, S.; Hua, W.; Ramanan, D.; Li, Y.; and Kong, S. 2023. Long-tailed 3d detection via 2d late fusion. *arXiv preprint arXiv:2312.10986*.

Mao, J.; Shi, S.; Wang, X.; and Li, H. 2023. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8): 1909–1963.

Peri, N.; Dave, A.; Ramanan, D.; and Kong, S. 2023. Towards long-tailed 3d detection. In *Conference on Robot Learning*, 1904–1915. PMLR.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Qian, R.; Lai, X.; and Li, X. 2022. 3D object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shi, G.; Li, R.; and Ma, C. 2022. Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *European Conference on Computer Vision*, 35–52. Springer.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770–779.

Shi, W.; and Rajkumar, R. 2020. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1711–1719.

Tian, X.; Ran, H.; Wang, Y.; and Zhao, H. 2023. Geomae: Masked geometric target prediction for self-supervised point cloud pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13570–13580.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, 11784–11793.

Zhan, J.; Liu, T.; Li, R.; Zhang, J.; Zhang, Z.; and Chen, Y. 2023. Real-aug: Realistic scene synthesis for lidar augmentation in 3d object detection. *arXiv preprint arXiv:2305.12853*.

Zhang, D.; Zheng, Z.; Niu, H.; Wang, X.; and Liu, X. 2023. Fully sparse transformer 3D detector for LiDAR point cloud. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4490–4499.

Zhu, B.; Jiang, Z.; Zhou, X.; Li, Z.; and Yu, G. 2019. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*.