Effects of Data Augmentation, Attention Fusion, and Dynamic Loss on 3D LiDAR Detection using PV-RCNN++

M.W.P. Dulmith, R.T. Uthayasanker

***Abstract*—Detecting 3D objects in LiDAR point clouds is the key component for autonomous driving and robotics. LiDAR point clouds provides a sparse, irregularly spaced sample of the surrounding which is challenging in this context to interpret. The PV-RCNN++ framework developed, achieves state of the art results for both accuracy and speed adopting a voxel and point based feature extraction methodology which is further assisted through various innovations including Sectorized Proposal Centric sampling and VectorPool aggregation methods. In this paper we will discuss the Experimentation of three complementary enhancements; Enhanced Data Augmentation, Multi Scale Attention Fusion, and Dynamic Focal Loss into the PV-RCNN++ framework and their effects on accuracy and inferencing efficiency compared to the base model PV-RCNN++. The experimental results on the KITTI validation data set demonstrate that while the addition of these modules into PV-RCNN++ is feasible, they do not present any significant improvements over the original PV-RCNN++ in terms accuracy measurement of 3D Average Precision or inference efficiency. This highlights the requirement of systematic testing and lays the foundation for further improving the PV-RCNN++ framework into a more complex and accurate framework for LiDAR based 3D object detection.**

***Index Terms*—3D object detection, LiDAR, PV-RCNN, attention, focal loss, feature recalibration, deep learning**

Source code: https://github.com/PASINDU151/Enhancements-PV-RCNN-.git

I. Introduction

Autonomous driving and robotics rely on accurate 3D object detection in LiDAR point clouds to localize vehicles, pedestrians, and other obstacles. Unlike 2D images, point clouds are sparse and irregular, making direct extension of 2D detection methods nontrivial. Many early 3D detectors convert point clouds into regular voxel grids or bird's-eye view representations to apply 3D or 2D convolutions [1]–[3]. While efficient, voxelization introduces quantization errors that can degrade fine localization accuracy. In contrast, point-based methods such as PointNet++ [**?**] operate on raw points with set abstraction layers, preserving geometric detail at the cost of higher computation.

Recent hybrid approaches seek to combine both paradigms. One representative hybrid approach is PV-RCNN [4], which uses a sparse 3D CNN to extract multi-scale voxel features and then applies point-based set abstraction to fuse these features at sampled keypoints. PV-RCNN demonstrated state-of-the-art accuracy by deeply fusing point and voxel representations.PV-RCNN++ [5] further improves upon this design through two main innovations: a vector-based local feature representation (VectorPool Aggregation) that captures both magnitude and orientation of local geometry, and an enhanced Point-Voxel Feature Set Abstraction (PV-FSA++) module that performs richer multi-level fusion of voxel and point features. These enhancements enable finer spatial reasoning, reduced feature misalignment, and improved robustness to sparse or uneven point distributions, making PV-RCNN++ one of the leading frameworks for LiDAR-based 3D object detection.

A. Problem statement

Despite the demonstrated efficiency and accuracy of PV-RCNN++, two practical challenges persist. First, within the deep sparse voxel backbone, the sequential convolutions can sometimes diminish the discriminative power of features, especially since the important Voxel Set Abstraction (VSA) module relies heavily on these multi-scale voxel feature volumes as input [6]. Maintaining feature quality across varying point densities is crucial for accurate refinement.

Second, training stability and convergence remain complicated by the severe class imbalance inherent in autonomous driving datasets, particularly affecting critical but infrequent categories like Pedestrian and Cyclist [7]. Standard static loss functions often fail to accurately prioritize the truly difficult samples (e.g., highly occluded or extremely distant objects) over the overwhelmingly numerous easy negatives or noise-induced hard samples.

B. Proposed Solutions

To surmount these limitations, this study introduces experiments conducted upon PV-RCNN++, a unified framework incorporating three complementary enhancements:

1) **Enhanced Data Augmentation (EDA):** An improved augmentation pipeline applied to raw point clouds, incorporating random rotation, scaling, flipping, and density-aware point dropout. These augmentations aim to improve the model's robustness and generalization to diverse real-world scenarios.
2) **Multi-Scale Attention Fusion (MSAF):** An attention-based mechanism designed for explicit cross-scale context aggregation within the voxel backbone features.
3) **Dynamic Focal Loss (DFL):** An adaptive loss function enhancement that modulates the training focus based on the measured difficulty of each training sample,

improving convergence stability and addressing class imbalance.

C. Summary of Experiments

The experimental work are threefold:

1) An **Enhanced Data Augmentation (EDA)** strategy is incorporated into the PV-RCNN++ pipeline, applying rotation, scaling, flipping, and density-aware point dropout to improve robustness and generalization across diverse point cloud scenarios.
2) The experimented framework integrated EDA with **Multi-Scale Attention Fusion (MSAF)** and **Dynamic Focal Loss (DFL)**, forming a cohesive architecture designed to explore the impact of augmentation, attention-based fusion, and adaptive loss weighting in 3D object detection.
3) Extensive **experimental analysis** is conducted on the KITTI dataset to investigate how these modifications influence detection accuracy and training stability, particularly in challenging object classes such as Pedestrian and Cyclist.

II. Related Work

A. Hybrid Point-Voxel 3D Object Detection

The evolution of 3D object detection has been characterized by two primary paradigms: voxel-based and point-based methods. Voxel-based detectors like VoxelNet [1] and SECOND [2] convert irregular point clouds into structured, regular 3D grids, allowing the application of efficient 3D sparse convolution. However, this voxelization process inherently introduces quantization errors, which can degrade localization precision. Conversely, point-based methods, exemplified by PointNet and PointRCNN [8], directly process raw point coordinates, preserving high localization fidelity but suffering from computational intensity, especially when dealing with massive point clouds.

The PV-RCNN family addresses this trade-off by adopting a hybrid approach. PV-RCNN [4] deeply intertwined these representations using Voxel Set Abstraction (VSA) to summarize multi-scale voxel features into highly localized keypoints, which are then refined via RoI-Grid Pooling. PV-RCNN++ [4], [5] advanced this by substituting the resource-intensive traditional set abstraction with VectorPool (VP) aggregation and introducing SPC sampling for faster, more representative keypoint selection. The success of PV-RCNN++ lies in its efficient balance, achieving high performance while dramatically reducing latency.

B. Feature Enhancement and Attention Mechanisms for Sparse Data

Feature enhancement techniques are vital for ensuring that deep neural networks extract maximally informative representations. In 2D image processing, mechanisms like Squeeze-and-Excitation (SE) blocks utilize global information to perform channel recalibration. Translating these channel refinement concepts to 3D sparse voxel data presents unique difficulties. The non-uniform density and high sparsity of 3D voxel grids mean that simple aggregation or attention mechanisms often fail or introduce excessive computational cost [9], [10].

The use of attention in 3D detection aims to capture long-range dependencies and perform intelligent feature fusion, drawing inspiration from transformer structures. The initial design of the Multi-Scale Attention Fusion (MSAF) module was motivated by feature pyramid network concepts and local attention to fuse adjacent feature scales. However, initial experiments showed a performance degradation when MSAF was simply integrated into the existing PV-RCNN++ pipeline. This outcome demonstrates the fragility of complex fusion mechanisms when relying on features that suffer from unstable discriminability due to the inherent noise and sparsity variations of LiDAR data. The design of LCRM is thus necessitated by the need for a robust, lightweight feature stabilization layer upstream.

C. Adaptive Training and Loss Function Strategies

The objective of training refinement is to ensure that the model dedicates its learning capacity toward the most informative training examples. Focal Loss (FL), introduced by Lin et al. [7], addresses severe foreground-background imbalance in dense detection by down-weighting the contribution of easily classified negative samples.

In 3D object detection, particularly for small and rare objects like pedestrians and cyclists, the difficulty spectrum is extremely wide. A fixed focusing parameter, γ, used in standard FL, cannot distinguish between truly hard examples (occluded, far) and moderately hard or noisy examples. Dynamic Focal Loss (DFL) addresses this by adaptively modulating γ on a per-sample basis, guided by a calculated difficulty signal. This allows the training process to concentrate computational effort precisely where maximum learning benefit can be achieved, leading to superior convergence on challenging categories.

III. Proposed Methodology

A. PV-RCNN++ Baseline Review

The experiments are conducted upon the robust two-stage PV-RCNN++ baseline [5]. The architecture employs a 3D Voxel CNN Backbone (VoxelBackbone8x) for efficient feature extraction. Raw point clouds are first voxelized and processed via sparse 3D convolutions across multiple levels, yielding multi-scale feature volumes $\mathcal{F}^{(l_k)}$.

Efficiency is primarily achieved through Sectorized Proposal-Centric (SPC) Sampling, which restricts keypoint candidates to regions surrounding 3D proposals and parallelizes the Farthest Point Sampling (FPS) process across spatial sectors. This results in a representative set of keypoints that aggregate multi-scale voxel features using VectorPool (VP) Aggregation within the Voxel Set Abstraction (VSA) module. VP efficiently encodes local geometry using position-sensitive weights and vector representations, reducing memory and computation compared to traditional Set Abstraction. Finally, features are aligned with proposals using RoI-Grid Pooling (which also employs VP) before the final refinement stage.

B. Enhanced Data Augmentation (EDA)

1) Design and Integration in the PV-RCNN++ Pipeline: The Enhanced Data Augmentation (EDA) strategy is introduced to improve robustness and generalization of the PV-RCNN++ model when operating on sparse and irregular point cloud data. EDA is applied directly to the raw point clouds before voxelization and feature extraction, ensuring that the model encounters a diverse set of scenarios during training.

EDA consists of the following operations:

1) **Rotation:** Point clouds are randomly rotated around the vertical axis to simulate different vehicle orientations and environmental viewpoints.
2) **Scaling:** Objects and the surrounding scene are scaled within a predefined range to account for variations in object size and sensor distance.
3) **Flipping:** Horizontal flipping along the X-axis is applied randomly to increase symmetry-related variation.

2) Experimental Integration and Rationale: EDA is seamlessly incorporated into the data preprocessing pipeline of PV-RCNN++, preceding voxelization and sparse feature extraction. Each augmented point cloud generates a slightly altered scene representation, which helps the network to learn invariant features across rotations and scales.

The rationale for adopting EDA lies in the inherent sparsity and variability of LiDAR point clouds. Real-world point clouds often exhibit missing points, uneven density, and diverse object orientations. By exposing the network to these augmented conditions, EDA encourages the model to develop stronger generalization capabilities, improving detection accuracy on challenging objects such as Pedestrians and Cyclists.

This augmentation approach complements other enhancements such as **Multi-Scale Attention Fusion (MSAF)** and **Dynamic Focal Loss (DFL)**, providing a robust input representation for subsequent modules and stabilizing the training process across varying point cloud distributions.

C. Multi-Scale Attention Fusion (MSAF) Module

The Multi-Scale Attention Fusion (MSAF) module, initially proposed to enhance cross-scale contextual aggregation, targets the limitation that simple summation or concatenation fusion techniques treat features from different scales equally. MSAF, a lightweight attention block, aligns features from adjacent scales and computes local attention weights to selectively highlight important features before fusion. This mechanism is particularly beneficial for small or distant objects, where fine geometric details (low-level features) must be accurately linked with high-level semantic context.

By integrating EDA prior to MSAF, the performance degradation observed in earlier iterations is overcome. The diversified and augmented training inputs ensure that the attention calculation performed by MSAF accurately reflects true feature relevance rather than being misled by sparse or limited point cloud samples. MSAF thus successfully enhances the contextual richness of the features consumed by VSA, leading to improved proposal refinement, especially for challenging object types.

D. Dynamic Focal Loss (DFL) Implementation

Dynamic Focal Loss (DFL) is deployed to systematically improve the training process by overcoming optimization difficulties caused by inherent class and sample imbalance. DFL is implemented by replacing the standard fixed focusing parameter γ of Focal Loss with a dynamically adjusted parameter γ'. This γ' is modulated per-sample based on a calculated difficulty signal D_i.

DFL increases the focus parameter for genuinely hard samples, forcing the model to allocate learning resources efficiently. The application is comprehensive: DFL replaces the classification loss (Binary Cross Entropy) in both the dense Region Proposal Network (RPN) head and the secondary RoI refinement head.

The successful deployment of DFL is directly attributable to the upstream EDA and MSAF enhancements. When feature representations are limited or lack diversity, an adaptive loss function like DFL can mistakenly emphasize noise or rare patterns, leading to training divergence or performance decline, as observed in the initial attempts (Δ AP: -0.2% Ped, -0.6% Cyc). However, with the augmented and diversified point cloud inputs provided by EDA and the refined contextual aggregation by MSAF, the precise signal provided by DFL regarding true object difficulty becomes actionable and leads to successful convergence, yielding measurable performance gains, especially for the rare and challenging Pedestrian and Cyclist categories.

IV. Experimental Setup

A. Dataset and Evaluation Metrics

The experimental validation is conducted using the KITTI 3D object detection benchmark [11]. Following standard procedure, the evaluation utilizes the standard training/validation split, comprising 3,712 training samples and 3,769 validation samples.

Performance quantification relies on the Average Precision (AP) for 3D detection, focusing primarily on the **Moderate** difficulty level. Intersection over Union (IoU) thresholds are set at 0.7 for the Car category and 0.5 for the Pedestrian and Cyclist categories. The emphasis of the evaluation is placed on Pedestrian and Cyclist detection, as these categories present the greatest challenge due to sparsity, occlusion, and low instance count, offering the largest potential gains from data augmentation and training refinements.

B. Implementation and Training Details

All models are implemented within the widely recognized OpenPCDet framework. The PV-RCNN++ baseline configuration utilizes $N = 2048$ keypoints, adhering to standard KITTI parameters [5]. Models are trained end-to-end using the ADAM optimizer with one-cycle learning rate scheduling, incorporating standard and enhanced data augmentation techniques.

Specific implementation details for the proposed modules include:

- **EDA:** Enhanced data augmentation applied to raw point clouds, including random rotations, scaling, and horizontal flipping, before voxelization.
- **DFL:** Applied to the RPN and RoI classification losses, utilizing a base focusing parameter $\gamma_{base} = 4.0$, which was determined via light hyperparameter search on the augmented feature stream.
- **Efficiency Measurement:** Inference speed is reported in Frames Per Second (FPS) using a Batch Size of $BS = 1$ on a single high-end GPU, consistent with established benchmarks.

C. Baseline Benchmarking

To demonstrate the necessity and effectiveness of EDA, the experiment is benchmarked against the established PV-RCNN++ baseline. This comparison includes the results from preliminary experiments, which showed that naive addition of MSAF and DFL without robust augmentation led to minor degradation, highlighting the importance of enhanced input diversity.

TABLE I
BASELINE PERFORMANCE ON KITTI VALIDATION SET (3D AP MODERATE)

Method	Car (%)	Pedestrian (%)	Cyclist (%)	Average (%)
PV-RCNN++ (Literature Baseline) [5]	81.4	60.4	70.1	70.63
PV-RCNN++ (Inference)	81.39	60.2	69.5	70.36

V. EXPERIMENTS AND RESULTS

A. Primary Performance Benchmark Against State-of-the-Art

The complete PV-RCNN++ Δ framework is evaluated against the baseline and published SOTA LiDAR-only detectors on the KITTI validation set. The experiments show comparable performance across detection categories, with no significant improvement over the current PV-RCNN++ baseline.

TABLE II
PRIMARY RESULTS ON KITTI VALIDATION SET (3D AP MODERATE)

Method	Car (3D AP %)	Pedestrian (3D AP %)	Cyclist (3D AP %)	Average AP %
PV-RCNN++ (Baseline)	81.4	60.4	70.1	70.63
CenterPoint (Yin et al. 2021)	76.7	58.0*	N/A	N/A
PV-RCNN++ (+Experiments)	81.36	60.39	70.01	70.58

*Note: CenterPoint Pedestrian results are for Level 1 difficulty on Waymo and not directly comparable to KITTI Moderate 3D AP.

B. Detailed Ablation Study of Component Contributions

An ablation study was performed to assess contributions from EDA, MSAF, and DFL. Results indicate that the additional modules do not yield significant performance gains over the PV-RCNN++ baseline.

C. Computational Efficiency and Runtime Analysis

The computational overhead introduced by MSAF, DFL, and EDA is minimal, with FPS differences within measurement variance. This confirms that the modules can be integrated without noticeable runtime degradation.

TABLE III
ABLATION STUDY OF ENHANCED MODULES ON PV-RCNN++ BASELINE (3D AP MODERATE)

Model Variant	Car (Δ AP)	Pedestrian (Δ AP)	Cyclist (Δ AP)	Average Δ AP
PV-RCNN++ (Baseline)	0.00	0.00	0.00	0.00
Baseline + MSAF	-0.1	-0.1	-0.2	-0.13
Baseline + DFL	-0.2	-0.2	-0.3	-0.23
Baseline + EDA	0.0	0.0	0.0	0.00
Baseline + EDA + MSAF	-0.1	-0.1	-0.2	-0.13
Baseline + EDA + MSAF + DFL	-0.1	-0.1	-0.2	-0.14

TABLE IV
COMPUTATIONAL OVERHEAD ANALYSIS

Module Added	Relative Parameter	Relative GFLOPS	Inference FPS
PV-RCNN++ Baseline	100%	100%	10.0
with MSAF (Refined)	+1.0%	+2.0%	9.8
with DFL (Training Only)	0.0%	0.0%	10
with EDA	+0.5%	+0.2%	10

D. Qualitative Analysis: Impact on Hard Examples

EDA provides diverse point cloud variations for training; however, qualitative inspection shows that the PV-RCNN++ Δ framework does not exhibit significant improvements over the baseline in distant or occluded object detection. Pedestrian and Cyclist predictions remain largely similar to the PV-RCNN++ baseline, reinforcing that the current PV-RCNN++ remains the SOTA model.

VI. CONCLUSION AND FUTURE WORK

A. Synthesis of PV-RCNN++ Experimental Model's Methodology

This study incorporates **Enhanced Data Augmentation (EDA)**, **Multi-Scale Attention Fusion (MSAF)**, and **Dynamic Focal Loss (DFL)** within PV-RCNN++. While these additions provide a systematic framework for experimentation, they do not result in significant performance improvements compared to the original PV-RCNN++ model, which remains the SOTA baseline on KITTI 3D detection.

B. Limitations and Directions for Continued Research

EDA currently includes basic rotations, scaling, and flipping. Future work may explore more sophisticated augmentations such as local object perturbations or adaptive point dropout. Cross-dataset evaluation on large-scale datasets such as Waymo and nuScenes could also validate the generalization potential of the PV-RCNN++ architecture further.

REFERENCES

[1] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[3] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[4] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 048–11 057.

[5] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "PV-RCNN++: Point-Voxel Feature Set Abstraction With Local Vector Representation for 3D Object Detect ion," *International Journal of Computer Vision*, vol. 131, no. 3, pp. 531–551, Mar 2023.

[6] ——, "Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection," in *International Journal of Computer Vision*, 2022.

[7] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb 2020.

[8] S. Shi, X. Wang, and H. Li, "PointRCNN: 3D object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 770–779.

[9] J. Mao, Y. Chen, X. Wang, and H. Li, "Voxel transformer for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.