AMP Final Assignment - Improving 3D Object Detection with Modular Enhancements to the CenterPoint Framework

Jasper Welgemoed Ishita Agarwal Taneshwar Pranav Parankusam Leander Le Ba

Abstract

This project investigates improvements to a CenterPoint-based LiDAR 3D object detector on the View of Delft dataset. Key extensions include semantic fusion via PointPainting, data augmentation for both LiDAR and image modalities, and architectural changes such as a multiview fusion neck and dropout. PointPainting, which enriches LiDAR points with RGB semantics through early fusion, proved most effective. Augmentation further improved robustness, while architectural changes offered limited gains under compute constraints. The best-performing model combined PointPainting and data augmentation, highlighting the value of semantic fusion and input-level strategies over more complex modifications in this setting.

1. Introduction

3D object detection from LiDAR point clouds is a cornerstone of autonomous driving systems, enabling vehicles to perceive and understand their surroundings with high spatial accuracy. By providing precise localization and dimension estimation, this capability is essential for tasks such as navigation, path planning, and collision avoidance. Among current approaches, voxel-based detectors, such as CenterPoint [15], achieve strong performance by converting sparse point clouds into structured grids, enabling efficient processing with 2D convolutional networks.

1.1. Problem Statement

Despite strong overall performance, voxel-based 3D object detectors such as CenterPoint often show reduced accuracy for smaller or less frequently occurring object categories. An initial evaluation on the View-of-Delft (VoD) dataset highlights this issue: while the baseline CenterPoint model detects cars reliably, it struggles with pedestrians and cyclists, classes that are both more difficult to detect and highly relevant for safety in urban environments.

An analysis of the standard CenterPoint configuration, which combines the SECOND backbone and SECONDFPN neck, reveals several contributing factors to this performance gap: Deep backbone networks reduce feature map resolution, which suppresses the fine details needed to detect small objects. Sparse point clouds and coarse voxelization reduce the representation quality of objects with fewer LiDAR points. Shared detection heads may not be sufficiently specialized to capture subtle differences in small or complex object shapes. Imbalanced training data with fewer examples of small object classes leads to weaker learning signals. To address these challenges, this work builds on the CenterPoint framework and investigates how performance can be improved through targeted modifications in data fusion, data augmentation, and network architecture, all within a constrained computational setting.

2. Related work

Recent progress in 3D object detection has been largely driven by LiDAR-based methods. Voxel-based approaches like SECOND [13] and CenterPoint [15] convert sparse point clouds into structured grids, enabling efficient feature extraction using 2D convolutions. These methods are widely adopted due to their balance of accuracy and scalability.

Sensor fusion is a common strategy for improving detection of small or occluded objects by combining LiDAR geometry with image semantics. Early fusion methods like PointPainting [12] augment point clouds with image-based class scores, while mid-level approaches such as BEVFusion [7] combine features in a shared BEV space, offering richer representations at higher computational cost.

Architectural improvements have also been explored. Deeper backbones like ResNet [3] help capture fine-grained features, while FPNs [6] and BiFPNs [10] enhance multiscale aggregation. Gating mechanisms [9] further refine feature fusion by learning to emphasize informative channels.

In addition, data augmentation plays a key role in improving generalization. As shown in [1], geometric and semantic-aware augmentations can mitigate class imbalance and improve robustness to domain shifts.

This work builds on these findings and addresses known limitations in the CenterPoint baseline by exploring targeted enhancements in fusion, augmentation, and network components under resource constraints.

3. Methodology

This section outlines the practical improvements made to the baseline CenterPoint detector for the View of Delft dataset, focusing on sensor fusion, data augmentation, and targeted architectural changes.

Given computational constraints, we avoided model assembly and augmentation in test time due to the added complexity [4] of managing multiple models and the limited applicability in the real world of the increased augmentation cost [8]. The encoder was left unchanged, aligning with recent work that emphasizes improvements in detection heads and data representation [2]. Hyperparameter tuning was done incrementally during development to balance performance gains with resource efficiency.

3.1. Sensor Fusion

Sensor fusion significantly enhances 3D object detection, particularly for distant, occluded objects or in complex scenes. Incorporating semantic information from RGB images can improve both classification accuracy and localization robustness.

This project uses only LiDAR and camera data. While radar can provide velocity information, its sparse and low-resolution output makes integration challenging [14]. Fusion strategies involving radar are also less established in the literature, making them harder to apply within a limited development timeline.

Early fusion was selected for its simplicity, modularity, and compatibility with the existing CenterPoint pipeline. Specifically, PointPainting [12] was implemented to enrich each LiDAR point with semantic class scores derived from RGB image segmentation. This approach allows semantic information to be incorporated directly into the input without altering the network architecture.

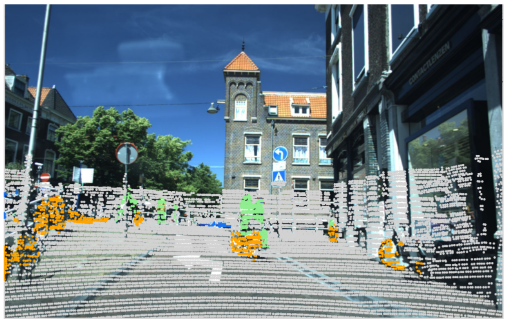

Figure 1. Painted LiDAR points projected onto the image plane, colored by semantic class: gray (background), orange (cyclist/bicycle), green (pedestrian), and blue (car). This illustrates the early fusion result using PointPainting.

The implementation follows a modular pipeline that integrates semantic features into the LiDAR point cloud. A DeepLabV3 model with a MobileNetV2 backbone is used for image segmentation, chosen over ResNet to reduce computation time while maintaining reasonable accuracy. The network is fine-tuned to predict four relevant classes: background, car, pedestrian, and cyclist. Using camera intrinsics and extrinsics, each LiDAR point is projected into the image plane, and the class probabilities of the nearest valid pixel are assigned to that point. These semantic scores are then appended to the XYZ coordinates, forming a 7D representation (excluding intensity), which is passed to a CenterPoint-based 3D detector without requiring architectural changes. A visual example of the painted LiDAR points projected onto the image plane is shown in Figure 1, where each point is colored according to its semantic label.

Additionally, BEVFusion [7] was considered. However, its training time exceeded the available compute budget by a significant margin, and it was therefore not pursued further.

3.2. Data Augmentation

To improve generalization and robustness, both image and point cloud augmentations are applied during training. Augmentation is performed after sensor fusion to maintain consistency with the painted point clouds.

Image augmentation affects 50% of the training samples, with no change in dataset size. For each selected sample, horizontal flipping, color jittering (brightness, contrast, saturation, hue), and grayscale conversion are applied independently with a 50% probability. When flipped, the LiDAR point cloud and ground truth boxes are mirrored accordingly to preserve alignment. *LiDAR augmentation* is applied on the 7D fused input after PointPainting. With a 50% chance, a sample is randomly rotated (±0.1 radians), scaled (±5%), and translated (up to ±0.2 meters) along all three axes. This introduces spatial variation while preserving semantic structure across modalities.

This augmentation strategy introduces useful diversity during training while preserving spatial and semantic consistency across modalities.

3.3. Backbone

To improve detection of smaller objects in 3D space, a ResNet backbone was introduced to the pipeline. ResNet's residual connections allow for effective training of deeper networks by mitigating the vanishing gradient problem. This depth enables the extraction of more expressive features, which is especially beneficial for recognizing pedestrians and cyclists, classes that are typically represented by fewer points and finer spatial details. By preserving these subtle patterns through deeper layers, ResNet offers a stronger foundation for detecting small or difficult objects compared to more lightweight convolutional backbones.

3.4. Neck

The neck module transforms multi-scale features from the backbone into a unified representation for the detection head. The baseline setup uses SECONDFPN, which upsamples and concatenates features from multiple layers of the SECOND backbone to form a high-resolution BEV map.

3.4.1 Single-Level Gated Multi-View Fusion

MultiViewFusionNeck was used, which combines information from the two complementary perspectives: Voxel Features: High resolution feature maps extracted from the initial stage of the SECOND backabone,preserving fine-grained spatial details for precise localization. Aggregated BEV features: High-level, context-rich features from concatenated SECONDFPN outputs, providing broader semantic information.

Using a gating mechanism that performs a learned, adaptive fusion of the voxel and BEV features: Both input feature maps are channel-aligned using 1x1 convolutions. A shared gating map *g* is computed using 1x1 convolution followed by a sigmoid activation function. This map captures spatial and channel wise importance weights. [9] The final fused feature map is computed as:

$$F_{\text{fused}} = v \cdot g + b \cdot (1 - g)$$

where: v is the reduced voxel feature, b is the reduced BEV feature, $g \in [0, 1]$ is the learned gate mask. A final 3×3 convolutional layer is applied to refine the fused output.

This gating-based approach was motivated by the understanding that different network stages encode complementary information (fine spatial detail vs. rich global context). Unlike traditional fusion strategies (e.g., concatenation or simple addition) that assume equal contribution, our method, inspired by prior work [5], allows the network to dynamically learn which features to emphasize.

3.4.2 Multi-Scale Gated Fusion

We attempted to re-architect SECONFPN into a BiFPN like structure to perform multi-scale gated fusion.

This modified design incorporates bidirectional flow throughout the feature pyramid. At each scale level, learnable *weightedfusion* blocks determined the relative importance of each feature map. These fusion block use trainable weights to adaptively combine inputs, allowing the network to emphasize the most informative features across resolutions.

BiFPNs are state-of-the-art for efficient multi-scale aggregation [11], enabling rich context exchange. However, this approach did not improve results, likely due to increased complexity and the need for more tuning or data.

3.4.3 Dropout Regularization

To enhance generalization and combat overfitting,*Dropout* was introduced. This technique randomly sets a fraction of feature map elements to zero during training, compelling the network to learn more robust features. Applied within SECONDFPN and MultiViewFusionNeck sequential blocks, a rate of 0.2 yielded promising results.

3.5. Head

Intermediate fusion techniques strike a balance between modularity, prallelism, and detection performance. In our early testing, we opted to combine LiDAR and Camera modalities to leverage their complementary strengths. To investigate this further, we adopted a BEVFusion-based approach [16], combining Aggregated BEV features from our SECONDFPN neck with projected image features in BEV Space. We used an image backbone network consisting of first two layers of ResNet18 to extract relevant feature images. Pre-trained image networks such as ResNet18 [3], as described earlier, effectively extract general features of known objects such as cars, pedestrians, and bicycles. To align the spatial dimensions of both feature sets, two seperate CNN's were employed, followed by a dedicated fusion network that produces multiple features maps. These are then passed on to the CenterPoint head. As mentioned in 3.1, the cost of implementing this fusion-based technique did not fit within the computational budget. In parallel, we explored hyperparameter tuning the CenterPoint head, which yielded lower performance than the baseline.

An overview of the complete architecture is provided in Figure 2.

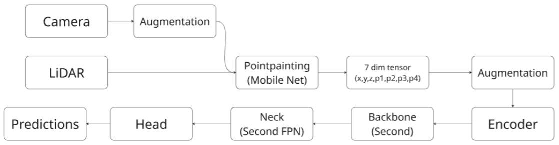

Figure 2. Full Pipeline Overview

4. Experiments

All experiments were conducted on the View of Delft (VoD) dataset, targeting 3D object detection for three categories: Cars, Pedestrians, and Cyclists. Detection performance was evaluated using the mean Average Precision (mAP) metric, computed per class based on Region of Interest (ROI) evaluation.

Each model was trained for up to 12 epochs using the AdamW optimizer with an initial learning rate of 0.001. A training batch size of 4 and a validation batch size of 1 were used. The same training protocol and data splits were maintained across all configurations to ensure fair comparison.

Table 1. Comparison of different configurations and their effect on 3D object detection mAP (Driving Corridor Area).

Method	**Car** mAP (%)	**Pedestrian** mAP (%)	**Cyclist** mAP (%)	**Overall** mAP (%)
Baseline (Standard FPN)	89.65	58.38	52.32	66.78
ResNet	70.00	54.00	73.00	66.00
Single-Level Gated Fusion	66.23	54.98	81.76	67.66
Multi-Scale Gated Fusion (BiFPN-like)	68.91	50.52	76.89	65.44
Gated Fusion + Dropout (0.2)	80.32	44.61	76.47	67.13
PointPainting (only)	70.65	66.88	85.59	74.37
PointPainting + Data Augmentation	70.47	72.52	85.55	**76.18**
Tuned CenterPoint Head	72.47	46.05	71.94	63.68

Table 2. Final performance of PointPainting + Data Augmentation + Gated Neck configuration

Class	**mAP (ROI)**
Car	90.17%
Pedestrian	73.95%
Cyclist	81.59%
Overall mAP	**81.90%**

Table 1 shows that the Single-Level Gated Fusion neck significantly improves Cyclist detection, although at the cost of reduced Car mAP, indicating a shift in focus toward smaller object classes. Introducing dropout partially recovers Car performance, while reducing Pedestrian mAP, suggesting possible over-regularization. The BiFPN-style multi-scale gated fusion does not provide clear advantages over the simpler single-level design.

On the data level, PointPainting yields strong gains across all classes, achieving a high overall mAP. The addition of data augmentation further increases Pedestrian mAP, resulting in an even better overall performance. The best-performing configuration then combines PointPainting, data augmentation, and the Gated Fusion neck, as summarized in Table 2.

5. Conclusion

This project investigated enhancements to the 3D object detection pipeline through architectural and data-centric modifications. Beginning with a baseline detector using a standard FPN neck, the study systematically evaluated the impact of different backbone architectures, fusion methods, and semantic augmentation techniques.

Experimental results showed that combining PointPainting with data augmentation significantly improved detection performance for vulnerable road users, especially pedestrians and cyclists, demonstrating the value of semantic information. Gated neck designs also contributed by better preserving discriminative features. In contrast, changes to the neck or backbone in isolation did not consistently surpass the baseline.

The final configuration, which is a combination of PointPainting with data augmentation and a gated fusion neck, achieved an overall mAP of 81.90 percent on the VoD test set, representing a clear improvement over the initial setup. These findings highlight the importance of integrating semantic enrichment and adaptive feature fusion to enhance 3D object detection accuracy.

Future work may explore optimizing the interaction between fusion strategies and semantic inputs, as well as incorporating temporal context to improve detection consistency in dynamic scenes.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers, 2022. 1

[2] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1201–1209, May 2021. 2

[3] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 1, 3

[4] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *arXiv preprint arXiv: 1611.10012*, 2017. 2

[5] Jaekyum Kim, Jaehyung Choi, Yechol Kim, Junho Koh, Chung Choo Chung, and Jun Won Choi. Robust camera lidar sensor fusion via deep gated information fusion network. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1620–1625, 2018. 3

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 1

[7] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multitask multi-sensor fusion with unified bird's-eye view representation, 2024. 1, 2

[8] Jongwook Son and Seokho Kang. Efficient improvement of classification accuracy via selective test-time augmentation. *Information Sciences*, 642:119148, 2023. 2

[9] Shounak Sural, Nishad Sahu, and Ragunathan Rajkumar. Contextualfusion: Context-based multi-sensor fusion for 3d object detection in adverse operating conditions, 2024. 1, 3

[10] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection, 2020. 1

[11] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection, 2020. 3

[12] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection, 2020. 1, 2

[13] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. 1

[14] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, and Yutao Yue. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Transactions on Intelligent Vehicles*, 9(1):2094–2128, Jan. 2024. 2

[15] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking, 2021. 1

[16] Hanqi Zhu, Jiajun Deng, Yu Zhang, Jianmin Ji, Qiuyu Mao, Houqiang Li, and Yanyong Zhang. Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion. *IEEE Transactions on Multimedia*, 25:5291–5304, 2023. 3