# Enhancing 3D Object Detection: A Multi-Modal BEV Fusion and Affine Augmentation Framework

Nicodemus Soh    Mukil Saravanan    Coen Mingelen    Rick de Graaf

## Abstract

*Relying solely on a single sensor, such as LiDAR, introduces inherent limitations in perceiving complete and reliable object states, particularly under conditions of occlusion, information sparsity, and sensor noise. To enhance the robustness and accuracy of 3D object detection for autonomous driving, we introduce two key techniques:* ***affine-transformation-based instance augmentation*** *and* ***Multi-Modal Bird's-Eye View (BEV) Fusion framework*** *encoding RGB images with LiDAR point clouds in the BEV feature space. Experimental results confirm that the instance augmentation alone significantly improves the detection accuracy by 7.11% mAP, while our late-fusion model (BEVFusion-L) achieved the highest detection performance of* ***+10.96% mAP*** *increase over the CenterPoint baseline model. Thus, these strategies collectively demonstrate the generalization performance of 3D object detection, particularly for small and occluded objects. The codebase is made available at* `https://github.com/Nicosoh/AMP_Final_Assignment/tree/working`

## 1. Introduction

Autonomous vehicles have the potential to make transportation safer, more efficient, and more accessible. However, before it can be safely deployed, it must be able to perceive and react to unpredictable environments. A key part of this is 3D object detection: the ability to locate and classify surrounding objects such as cars, pedestrians, and cyclists in three-dimensional space. Camera, RADAR, and LiDAR are the most common sensors used, and among these, LiDAR is often preferred because of its ability to generate precise 3D point clouds of the environment. Many state-of-the-art (SOTA) 3D object detectors, such as CenterPoint, rely mainly on LiDAR for object detection and classification [18]. To counter these limitations and improve the current CenterPoint baseline, we propose an approach that, next to LiDAR data, uses a secondary RGB image modality. BEVFusion combines data from both sensors to enhance detection performance. In addition to addressing the observed overfitting and poor pedestrian detection, we apply data augmentation to increase data diversity and overall generalization. While CenterPoint achieves SOTA performance on LiDAR-based 3D object detection benchmarks, it relies solely on a single modality. This reliance becomes a limitation in scenarios where the LiDAR data is sparse, noisy, or occluded.

This paper is structured as follows: Section 2 reviews related work. Section 3 describes our proposed methodology for BEVFusion and data augmentation. Section 4 presents our experimental setup, results, and analysis. Finally, section 5 concludes the paper.

## 2. Related work

### 2.1. Lidar-based 3D Object Detection

To place our approach in context within the landscape of LiDAR-only 3D detection, we review SOTA architectures. Influential LiDAR-only detectors include voxel-based VoxelNet [19], its sparse-convolution successor SECOND [17], the pillar-based PointPillars [6], the hybrid two-stage PV-RCNN [13], and the anchor-free CenterPoint [18]. VoxelNet divides the cloud into a 3D grid and applies dense 3D CNNs, resulting in strong accuracy at the expense of high computational cost. SECOND reduces this load via sparse convolutions, trading off some fine detail. PointPillars projects vertical columns into a BEV pseudo-image for fast 2D CNN processing, sacrificing vertical resolution for real-time speed. PV-RCNN generates coarse BEV proposals before refining them with a point-based head, achieving precise localization with minimal overhead. CenterPoint learns a BEV heatmap of object centers and directly regresses sizes, orientations, and velocities, an anchor-free, rotation-invariant design delivering SOTA single-LiDAR performance in a modular framework.

### 2.2. Multiply Modalities Fusion for 3D Detection

Fusion of complementary modalities like LiDAR and cameras is a popular approach that has been extensively explored. Vora et al. published a sequential fusion technique called PointPainting, which decorates LiDAR points with class information through semantic segmentation [15]. This technique increased the mAP by 6.3 on the nuScenes dataset and was impressive based on its simplicity [2]. Philion et. al proposed Lift, Splat, Shoot (LSS) for unprojecting camera features back into 3D space [12]. This allowed multi-view cameras to transform from 2D images into bird's eye view (BEV) through probabilistic estimation of depths and camera intrinsics. This laid the foundation for multiple versions of BEVFusion [9] [10]. This area of BEV-related approaches is currently still an active area of research, with fusion through attention mechanisms [7], combining multiple complimentary tasks such that the model can learn better representations [8] and also modality dropout to account for missing sensor input [16].

### 2.3. Overfitting and Generalization in 3D Object Detection

From the baseline Center-point model, it can be seen that the model is over-fitting after epoch 7 due to the diverging train and validation loss. It has 5.2M parameters with only 5139 train samples while for comparison, ResNet34 [4] with 21.8M parameters was trained on ImageNet [3] with 1.2M images which is 55 times the number of samples per parameter. There are various methods to augment data hence creating variability to reducing the probability of overfitting. Zhu et al. conducted a survey of augmentation techniques ranging from basic to specialized [20]. The most common augmentation techniques are the affine transformations, like rotation, flipping, scaling, and translation. A particular limitation of the VoD dataset is its annotation scope, restricted solely to the camera's field of view, unlike datasets offering full 360-degree object annotation. This spatial constraint likely contributes to a lower average number of instances per frame. Consequently, the combined challenge of limited training samples and sparse instance density per frame highlights the critical role of data augmentation in fostering robust model generalization.

## 3. Methodology

### 3.1. Overview

To mitigate the inherent limitations of the LiDAR-only CenterPoint baseline on the VoD dataset [11], notably its susceptibility to overfitting and suboptimal pedestrian detection, we propose two principal enhancements. First, a multi-modal fusion approach that integrates RGB camera data with LiDAR. Second, a data augmentation pipeline, encompassing affine transformations is applied to improve generalization and data diversity. These modular contributions are evaluated both individually and in combination to comprehensively assess their effect on 3D object detection performance.

### 3.2. Multi-Modal BEV Fusion

Our fusion method builds upon the CenterPoint framework [18], which processes LiDAR and image modalities to create a unified Bird's-Eye View (BEV) representation for robust 3D object detection. For the LiDAR pipeline, it follows Centerpoint, which builds upon the PointPillars framework for LiDAR feature extraction. For the image peipeline, it employ a ResNet50 [4] backbone truncated at layer 3 to extract deep semantic features from the front-facing RGB camera. Since these image features are in perspective view and not aligned with the LiDAR features, it is transformed into the BEV domain using LSS [12].

The BEV features from both modalities are concatenated channel-wise, forming a multi-channel BEV representation. Inspired by TransFusion, the BEV representation is then fed through 4 layers of 2d Convolutions, Batch Normalization, and ReLU activations instead of an attention mechanism [1]. The learned fusion module will perform 2 roles, firstly to correct possible spatial misalignment and second to filter out unnecessary features. The overview of the proposed model is shown in Figure 1.

Additionally, another version of BEVFusion, which we term BEVFusion-L, where L stands for late fusion. The image pipeline remains the same except that ResNet50 is swapped for ResNet34. The BEV representation of the LiDAR points first goes through SECOND and SECONDFPN and then is passed to the fusion module together with the LSS features from the image pipeline, and lastly followed by the head. The reasoning behind this is to fuse only after meaningful features from the pseudo images (from Pillar Scatter) have been extracted. *Note: BEVFusion-L was trained over 12hrs and hence is not considered for submission and only used in the analysis and discussion.*

### 3.3. Data Augmentation

To increase invariance in the training dataset, we adopt an instance-based augmentation technique. As shown in Table 1, the 5,139 training frames contain a total of 38,436 annotated objects. While the Pedestrian and Car classes are relatively well represented, Cyclists are significantly underrepresented. Moreover, most instances of the Cyclist and Pedestrian classes are either fully visible or only partially occluded, with just 17.4% and 13.1%, respectively, labeled as largely occluded.

Cars tend to appear at greater distances on average ($\mu = 31.73\,\text{m}$), while Pedestrians are typically closer ($\mu = 23.61\,\text{m}$). Cyclists fall in between ($\mu = 26.01\,\text{m}$), with all classes exhibiting comparable spatial variance.

These statistics highlight clear imbalances in both class frequency and spatial distribution, underscoring the need for augmentation strategies to improve model generalization. Inspired by He et al. [4], we leverage affine transformations, including random flipping along the $x$-axis, rotation within $\pm 0.7$ radians, scaling in the range $[0.95, 1.05]$, and translation with Gaussian noise of standard deviation 0.5 meters along each axis.

To further mitigate class imbalance, we created a ground truth database with the training data and implemented

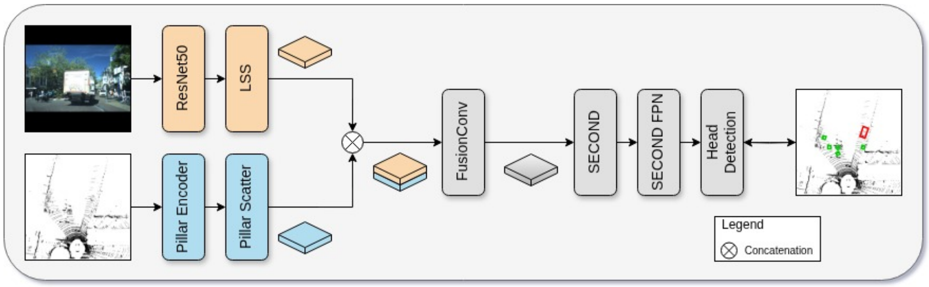

Figure 1. Overview of the BEVFusion architecture, which performs multi-sensor fusion in the BEV feature space.

ground truth sampling for both the LiDAR-only and the LiDAR/camera fusion settings. Although effective in diversifying training data, this method was ultimately excluded due to its increased computational requirements under our training constraints. The data augmentation pipeline was largely adapted from the OpenPCDet repository [14], with considerable modifications to accommodate our coding environment and the specific structure of the dataset.

| **Class** | **Count** | **Fully Visible** | **Partly Occluded** | **Largely Occluded** |
|---|---|---|---|---|
| Car | 15608 | 12.6% | 56.8% | 30.6% |
| Pedestrian | 16143 | 45.4% | 37.3% | 17.4% |
| Cyclist | 6685 | 68.4% | 18.5% | 13.1% |
| All Classes | 38436 | 36.1% | 41.9% | 22.0% |

Table 1. Object counts and occlusion percentages by class (total frames: 5139).

### 3.4. Training Details

Evaluation of the CenterPoint model with augmentation was conducted using a batch size of 16. Training adhered to the original CenterPoint training [18]. The model was trained for 14 epochs, with weights from epoch 11 selected for final evaluation. A learning rate of 0.001 was employed.

Our implementation of BEVFusion was trained using the largest batch possible, which was 6. This is because CenterPoint and one of the original BEVFusions were trained with a batch of 16 and 32, respectively, signifying the preference for larger batches [18] [9]. A maximum learning rate of 0.001 was used with cosine annealing, which smoothly varies the learning rate to promote training stability in the short training window of 4 hours, and the model was trained for a total of 8 epochs. For the image encoder, ResNet50 pretrained on ImageNet was utilized and frozen; the rest were either randomly initialized or with He Initialization for training stability with ReLU activations [5].

Lastly, for BEVFusion-L, a batch of 4 was used. It was trained for 18 epochs, and the weights from epoch 16 were used for the final model. The learning rate was set at 0.0001.

## 4. Experiments

### 4.1. Dataset and Experimental Setup

The View of Delft (VoD) dataset serves as the foundation for training and validating our framework. We focus on the detection of three target classes: **Cars, Pedestrians, and Cyclists**. Our proposed approach leverages two modalities: 3D LiDAR point cloud and RGB image zero-padded to a square and resized to 640 for BEVFusion and resized to 1280 for BEVFusion-L. Since data augmentation is done by applying affine transformations on instances, the total instances of the target class after data augmentation remains the same. The dataset is split into 70% for training and 30% for validation from its total of 7,386 frames.

The performance of the object detection models is quantified using the Mean Average Precision (mAP), which assesses both localization accuracy and classification performance by averaging the Average Precision (AP) across all object classes and multiple Intersection over Union (IoU) thresholds. A comparative analysis is primarily performed against CenterPoint. It is to be noted that, without explicit data augmentation, the baseline CenterPoint model achieves a mAP of 66.78 on the VoD test set, serving as the reference for evaluating subsequent methods.

### 4.2. Quantitative Results

Table 2 presents the results on the VOD unseen test set, all the variations of the proposed model outperform the CenterPoint, with CenterPoint with data augmentation slightly edging out BEVFusion with data augmentation. BEVFusion - L with data augmentation performed the best with +10.96 over the baseline.

| Method | Modality | mAP (%) |
|---|---|---|
| CenterPoint | L | 66.78 |
| CenterPoint + Data Aug | L | 73.89(+7.11) |
| BEVFusion + Data Aug | L + C | 73.05(+6.27) |
| BEVFusion - L + Data Aug | L + C | 77.74(+10.96) |

Table 2. Driving Corridor mAP on the VOD Test set

It is observed that BEVFusion-L with data augmentation strongly validates that fusing features after significant semantic and spatial information extraction from each modal-

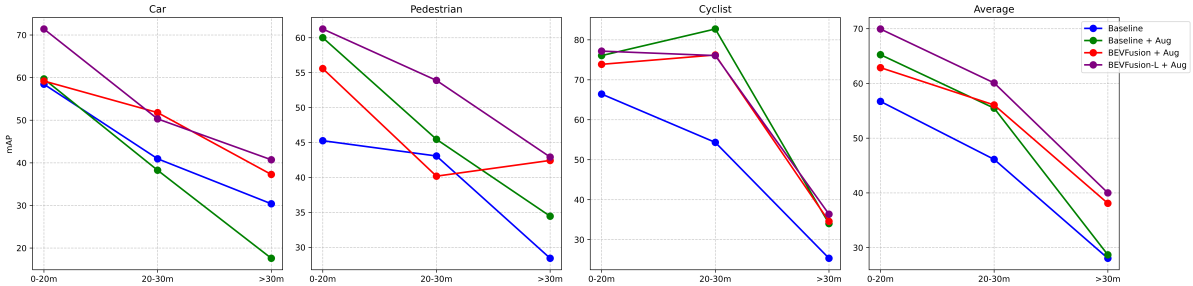

Figure 2. Distance vs Class vs mAP for each trained model

ity leads to more robust and accurate detections. The consistent mAP gains across various configurations when employing data augmentation confirm the hypothesis of limited dataset size and inherent class or occlusion imbalances in the baseline framework.

## 4.3. Discussion

**Impact of Data Augmentation:** Overfitting was evident in the baseline CenterPoint model, as observed from its divergence between its training and validation loss. Introduction of the augmentation to CenterPoint mitigated this occurrence and also positively impacted the model's performance by 7.11 mAP. From Table 1 it can also be seen that with augmentation, the baseline model outperforms in all target classes and distances except for cars at 20-30m and >30m. The cause of the drop in those categories is, however, unknown, as affine transformations do not modify the density but only the physical location and size. A deeper investigation is required to find the root cause. Notably, the performance improvement is also carried across to both BEVFusion models.

**Impact of Multi-modal Fusion:** Integrating both LiDAR (L) and Camera (C) modalities validates the complementary strengths of LiDAR's precise point cloud and camera's rich semantic features. BEVFusion with data augmentation outperformed CenterPoint, but underperformed CenterPoint with augmentation, even with an additional modality of the camera. Analysis was performed with visual inspection of the outputs from the LiDAR BEV stream and the camera BEV stream. The camera BEV stream displayed signs of high similarity between all channels, pointing to the possibility of a faulty implementation of LSS. However, this might also be due to the softmax on depth probabilities. Moving to BEVFusion-L, this outperforms CenterPoint with Augmentation, which suggests that the fusion between modalities should only be after both respective backbones. Future improvements could also be an additional BEV backbone followed by FPN after fusion for additional feature extraction before the head, and a dedicated FPN for the image stream.

From Figure 2 it is observed from the average mAP across distances, BEVFusion-L outperforms all other models, while CenterPoint with augmentation is on par with BEVFusion except for >30m. This could possibly mean that the camera stream is adding semantic information to ranges >30m. However, cameras are similar to LiDAR with increased sparsity/decreased resolution for further ranges. Looking at the individual classes, similar performance can be observed for cyclist, while for cars and pedestrians, equal or improved performance by BEVFusion is observed, except for pedestrians with BEVFusion at 0-20m and 20-30m. This is probably because, since pedestrians/cyclists are smaller objects as compared to cars, it does not provide additional information. However, cars are still visible at the 20-30m and >30m, hence improving performance and elevating the overall mAP.

## 5. Conclusion

This study addressed the limitations of LiDAR-centric 3D object detection, particularly overfitting and generalization challenges on the VoD dataset. The demonstrated data augmentation via instance-level affine transformations effectively mitigates overfitting with +7.11% mAP improvement from the CenterPoint baseline. Furthermore, our investigation into multi-modal BEV-fusion revealed that the late-fusion strategy enhances the mAP by +10.96% over the baseline, underscoring the advantages of fusing semantically rich representation for robust generalization. Future work will focus on refining augmentation techniques for small, distant objects and effectively using dense depth information from Image data.

## 6. Contributions

- N. Soh developed BEVFusion, BEVFusion-L & Affine Augmentations and wrote 2.2, 2.3, 3.4, 4.2, 4.3.
- M. Saravanan developed depth estimator & segmentation encoder with LiDAR fusion (not used eventually) and wrote Abstract, 4.1, 4.2, 4.3, 5.
- C. Mingelen developed GT database and sampling code for both LiDAR only model and LiDAR, image fusion (not used eventually) and wrote 3.3.
- R. de Graaf developed Dynamic Voxelization, PointPainting(not used eventually), Data processing and visualizations and wrote 1, 2.1, 3.1, 3.2.

## 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. 2

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015. 3

[6] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 1

[7] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan Yuille, and Mingxing Tan. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17182–17191, June 2022. 2

[8] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[9] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework, 2022. 2, 3

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

[11] Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian F. P. Kooij, and Dariu M. Gavrila. Multi-class road user detection with 3+1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7(2):4961–4968, 2022. 2

[12] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d, 2020. 2

[13] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection, 2021. 1

[14] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020. 3

[15] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[16] Shiming Wang, Holger Caesar, Liangliang Nan, and Julian F. P. Kooij. Unibev: Multi-modal 3d object detection with uniform bev encoders for robustness against missing sensor modalities, 2024. 2

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

[18] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking, 2021. 1, 2, 3

[19] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection, 2017. 1

[20] Qinfeng Zhu, Lei Fan, and Ningxin Weng. Advancements in point cloud data augmentation for deep learning: A survey. *Pattern Recognition*, 153:110532, Sept. 2024. 2