LiDAR-Assisted 3D Human Detection for Video Surveillance

Miquel Romero Blanch[1,2]
miquel.robla@gmail.com

Zenjie Li[2]
zli@milestone.dk

Sergio Escalera[1,3]
sescalera@ub.edu

Kamal Nasrollahi[2,3]
kn@create.aau.dk

Universitat de Barcelona and Computer Vision Center [1]

Milestone Systems[2]

Aalborg Universitet[3]

Abstract

This work explores 3D object detection using LiDAR technology, specifically focusing on pedestrian detection for video surveillance. While LiDAR is well-established in autonomous driving, its application in video surveillance is underexplored. We adapt state-of-the-art autonomous driving models for video surveillance, with CenterPoint being the top performer. Optimizing hyperparameters, such as voxel size and sweep merging, enhances pedestrian detection. Incorporating larger range data aids in generalization for video surveillance scenarios. This research demonstrates the feasibility of pedestrian detection in video surveillance and highlights open challenges related to domain adaptation and the high cost of high-resolution LiDAR sensors. Code: https://github.com/0Miquel/OpenPCDet-video-surveillance.

1. Introduction

In recent years, video surveillance has become essential for public safety and security. However, traditional 2D object detection has limitations in estimating depth and handling complex scenarios. To address these issues, LiDAR sensors, known for their precision, have gained attention in video surveillance. LiDAR sensors use lasers to create detailed 3D point clouds of the environment. While LiDAR is commonly associated with autonomous driving, its applications extend further. LiDAR enhances scene understanding and privacy preservation in video surveillance. Unlike cameras that capture identifiable images, LiDAR generates anonymous point clouds, addressing privacy concerns, particularly in public spaces.

The rise of LiDAR is driven by advances in autonomous driving. Most existing large-scale datasets for 3D object detection, like nuScenes [2] and Waymo [15], focus on autonomous driving scenarios. Given this limitation, our study explores state-of-the-art models originally designed for autonomous driving and applies them to video surveillance, with a specific emphasis on pedestrian detection.

This study presents the following contributions:

- **Evaluate Existing 3D Object Detection Models for Pedestrian Detection**: This objective involves an extensive evaluation and comparison of state-of-the-art models, including SECOND [16], PointPillar [7], and CenterPoint [18]. Performance metrics, such as accuracy and speed, will be considered.
- **3D Object Detection Evaluation for Video Surveillance**: The study aims to leverage data from autonomous driving scenarios to create an evaluation context that closely resembles video surveillance challenges, particularly in human detection. Insights drawn from this approach will inform the development of 3D object detection systems for real-world video surveillance.
- **Develop a 3D Human Detection Optimization Pipeline**: The study seeks to design and implement a dedicated pipeline for optimizing 3D object detection models, depicted in Fig. 1, specifically for human detection. This pipeline will encompass various hyperparameters, configurations, and data pre-processing techniques tailored to enhance accuracy and reliability in detecting humans.

2. Related work

In recent years, LiDAR-based 3D object detection has seen remarkable progress, transforming industries like autonomous driving, robotics, and video surveillance. This section provides a review of the state of the art in 3D object detection with LiDAR, with a focus on identifying techniques relevant to video surveillance applications.

LiDAR-based 3D object detection is notable for its ability to provide precise and dense spatial information, enabling accurate environmental perception. Within this domain, three main paradigms have emerged: point-based, voxel-based, and multimodal approaches.

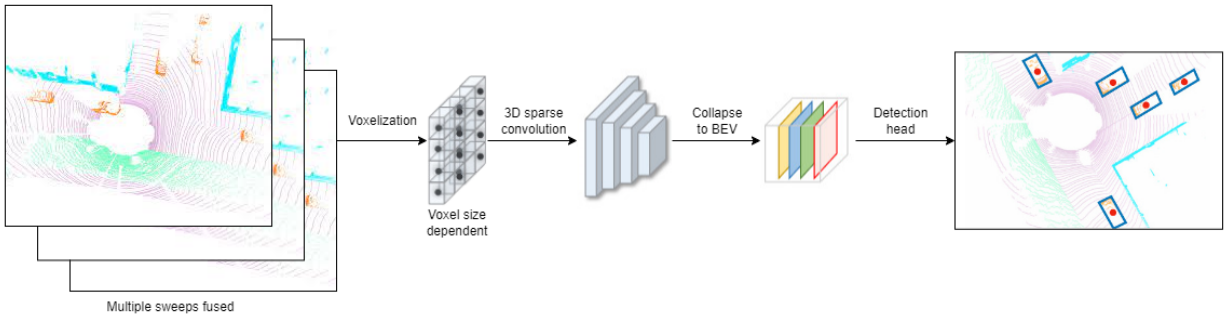


Figure 1. The presented optimized pipeline leverages state-of-the-art techniques to achieve peak performance in pedestrian detection. It is designed with a voxel-based approach, taking into account the challenges presented by the nuScenes dataset. Through this pipeline, we aim to comprehensively explore the impact of various hyperparameters and uncover the rationale behind our results.

2.1. Point Cloud 3D Object Detection

3D object detection using point clouds from LiDAR sensors can be summarized into two main categories: point-based methods and voxel-based methods.

2.1.1 Point-based methods

Point-based techniques, pioneered by PointNet [9], directly process unordered point cloud data. PointNet introduced a 3D deep learning architecture capable of handling tasks like object classification, segmentation, and later, object detection. PointRCNN [11] extended this approach with a two-stage framework involving proposal generation and proposal refinement in canonical coordinates to yield final detection results. While point-based methods maintain precise point location and offer flexible receptive fields, they often spend significant computation organizing irregular point data, making them less suitable for large-scale point clouds. Another emerging trend in point cloud processing involves the use of GNNs to learn permutation-invariant representations implicitly. GNNs are well-suited for point clouds, which can be naturally represented as graphs. Point-GNN [12], for instance, reasons on local neighborhood graphs constructed from point clouds, where each node iteratively summarizes semantic information from its neighboring points.

2.1.2 Voxel-based methods

Most existing methods nowadays tend to convert the point cloud data into a structured representation, typically a 3D voxel grid, to enable efficient processing. Voxelization provides a grid-based representation that allows the utilization of 3D convolutional neural networks (CNNs) for object detection. The first work that considered this approach was the pioneering VoxelNet [20], which divides the point cloud into 3D voxels, and encodes scene feature using 3D convolutions. However, its computational cost makes it difficult to use for real-time applications.

A more efficient approach was presented by SECOND [16], which proposed to use 3D sparse convolutions to tackle the high number empty voxels due to point cloud sparsity, improving VoxelNet results in term of both accuracy and speed. Therefore, SECOND has been a very well established baseline so far. On the other hand, PointPillars [7] [14] proposed an alternative voxelization approach by dividing the point cloud into pillar-shaped volumes, which reduced memory consumption and enabled efficient training and inference. It employed a lightweight 2D CNN for feature extraction, followed by a sparse convolutional network for object detection.

All methods proposed at that point used an anchor-based detection head. It was up until that moment when CenterPoint [18] introduced a center-based voxelization scheme that enhanced the efficiency and accuracy of 3D object detection. By representing objects as 3D center points, it achieved superior performance on various benchmarks while maintaining a low computational overhead.

At the same time, other works like PillarNet [10] opts to extend the concept of PointPillars and proposes a more powerful encoder network for effective pillar feature learning, a neck network for spatial-semantic feature fusion and the same detection head proposed by CenterPoint. As such, they design the first real-time and high-performance pillar-based 3D detection method.

2.2. Multi-modal 3D Object Detection

In the context of autonomous driving, LiDAR sensors usually come together with multiple camera sensors which provide RGB data of the whole scene. This has motivated the idea of using multi-modal methods that can perform a more robust estimation at the cost of losing the preservation

of privacy. Additionally, such approach has been proved to help detecting pedestrians as it provides denser information.

MVX-Net [13] proposed an early-fusion approach to combine the RGB and point cloud modalities, by leveraging the pioneering VoxelNet architecture and extracting the image features from a pretrained 2D Faster R-CNN.

One of the main issues that was identified was the association between LiDAR points and image pixels, which most of the times was challenging due to external factors like LiDAR malfunctions or inferior image conditions. BEVFusion [8] proposed a novel yet simple framework that unifies camera and LiDAR features in a shared BEV space instead of mapping one modality to the other, preserving camera's semantic density and LiDAR's geometric structure in order to address possible malfunctions. At the same time, TransFusion [1] also proposed a robust solution to LiDAR-camera fusion with a soft-association mechanism to handle inferior image conditions. Specifically, their implementation consists of a convolutional backbone and a detection head based on a transformer decoder.

3. Methodology

This section presents the datasets and models that will be considered. Additionally, a benchmarking protocol has been defined in the context of video surveillance.

3.1. Data

Video surveillance is essential for public safety and security, involving the monitoring of video streams to detect objects like pedestrians and vehicles, often in indoor spaces like shopping malls and parking lots. Unfortunately, there's a lack of indoor datasets with pedestrian annotations, which posed a challenge for our 3D object detection research. To overcome this limitation, we expanded our investigation to outdoor datasets primarily designed for autonomous driving.

Addtionally, LiDAR devices play a crucial role in our study due to their spatial and temporal resolutions, which impact object detection. Spatial resolution, determined by emitted beams, influences detail capture, while temporal resolution, determined by sweep frequency, affects data timeliness and dynamism. Our study examines these resolution aspects to align with our goal of creating efficient detection models tailored for video surveillance.

3.1.1 Outdoor datasets

The surge in autonomous driving has resulted in a wealth of outdoor datasets for 3D object detection using LiDAR. These datasets are rich in high-quality pedestrian annotations, making them invaluable for video surveillance applications.

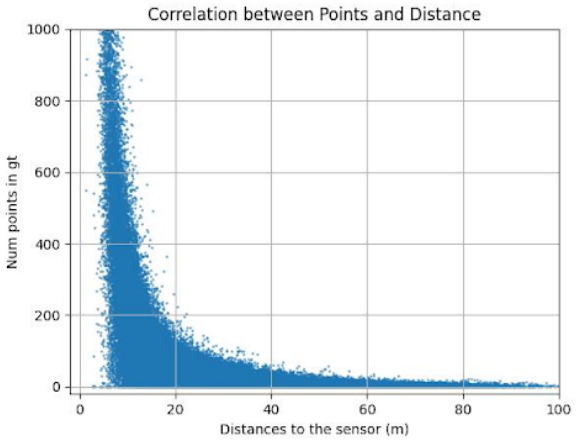


Figure 2. Correlation between points and distance to the sensor for pedestrian detections in nuScenes dataset.

In our search to identify the most fitting dataset for our video surveillance research, we meticulously assessed popular autonomous driving datasets listed in Tab. 1. Our evaluation considered crucial factors such as annotation quality, diversity, scene variety, and the inclusion of RGB images for multi-modal approaches.

After a comprehensive evaluation, the nuScenes dataset [2] emerged as the top choice for video surveillance. It offers meticulous annotations, including 3D bounding boxes, orientation, and velocity for pedestrians. Notably, nuScenes aligns closely with our intended LiDAR setup, featuring low spatial resolution (32 beams) but a high temporal resolution (20Hz capture frequency). This resemblance to our target configuration makes nuScenes an ideal choice for developing and testing models suitable for real-world deployment scenarios. To enhance its utility, we leverage multiple consecutive LiDAR sweeps, creating denser representations. These dense representations span up specifically to 20 meters, depicted in Fig. 2, a crucial range for video surveillance.

Furthermore, the cost-effectiveness of low spatial resolution LiDAR devices adds to nuScenes' appeal, aligning well with practical deployment considerations. While datasets like KITTI [5] and Waymo [15] are recognized, KITTI's scale is limited, and Waymo posed challenges due to licensing restrictions. The nuanced features of nuScenes, along with its convenient mini dataset for experimentation, solidified its position as our dataset of choice.

3.1.2 Indoor datasets

Initially, our focus was on indoor datasets for pedestrian detection. However, due to the scarcity of such datasets with pedestrian annotations, we had to turn to outdoor alterna-

Dataset	Year	Size						Diversity		
		Train	Val	Test	3D boxes	PC frames	RGB imgs	Scenes	Classes	Night/Rain
KITTI [5]	2012	7k	-	7k	200K	15K	15k	50	3	No
Argoverse [3]	2019	275k	105k	87k	993K	44k	490k	113	15	Yes
Lyft L5 [6]	2019	136k	-	164k	1.3M	46k	46k	366	9	No
nuScenes [2]	2019	168k	36k	36k	1.4M	400k (40k)*	1.4M	1000	10	Yes
Waymo [15]	2020	611k	152k	200k	112M	200k	1M	1150	3	Yes

Table 1. Statistics from the datasets studied. *nuScenes only has 40k annotated frames from the 400k total frames available.

tives. Still, we recognize the importance of indoor scenes in video surveillance.

Indoor datasets like SUN-RGBD [14] and ScanNet [4] are valuable but lack the specific pedestrian annotations needed. Instead, the L-CAS dataset from the University of Lincoln [17] fits our requirements better. This indoor dataset uses a LiDAR device with 16 beams and closely simulates real-world indoor surveillance scenarios.

However, the L-CAS dataset has a limitation: it lacks annotations for many pedestrians, especially those at a distance.

3.1.3 Other Point Cloud Capture Systems

Beyond LiDAR, there exist alternative point cloud capture systems, including structured light scanners , stereo cameras and radar. Structured light scanners, while offering accurate 3D object shape capture, are constrained by limited working ranges, rendering them unsuitable for large surveillance areas. Similarly, stereo cameras suffer from restricted field views and computational demands, hindering their efficacy in comprehensive video surveillance applications. Radar, while applicable in certain scenarios, faces issues with shape accuracy and resolution as their working ranges can be very limiting.

Considering these factors and the predominant focus of available datasets on LiDAR data, it is evident that LiDAR devices offer the most promising path for 3D object detection.

3.2. Model

Voxelization methods have emerged as a popular choice in 3D object detection tasks, as studied in the state of the art section, due to their ability to efficiently represent 3D data using regular grids of voxels. These methods provide a natural extension from 2D convolutional neural networks (CNNs) to 3D CNNs. By discretizing the 3D space into voxels and encoding the data in a structured manner, voxel-based models efficiently process 3D point clouds while benefiting from the well-established and optimized 3D CNN operations. This results in impressive accuracy without compromising inference speed thanks to sparse convolutions, making them well-suited for real-time applications like pedestrian detection in video surveillance.

Apart from voxelization, pillar encoding, which organizes point clouds into vertical pillars, is being investigated as an efficient alternative. Pillar-based models operate directly in 2D without 3D backbones.

To optimize pedestrian detection, two detection heads are considered. The anchor-based method, traditional but computationally intensive, contrasts with the center-based approach, newer and more efficient, directly predicting object centers in BEV space. This exploration seeks to determine the most effective approach for accurate and efficient pedestrian detection in diverse indoor and outdoor scenarios.

As a result, for our pedestrian detection experiments, we have chosen four models from the studied state of the art: SECOND [16], CenterPoint [18], CenterPoint with Pillar Encoding, and PillarNet [10]. These models use a consistent point cloud encoding scheme, including dimensions like x, y, z coordinates, and intensity. The inclusion of intensity is crucial for object detection, providing insights into the reflectivity of laser pulses. Additionally, a timestamp dimension is added to accommodate multiple sweeps for each sample, aiding in processing the temporal aspect of the scene and improving object dynamics and movement capture. With such temporal dimension it is expected to inherently estimate the motion compensation.

In summary, our model research encompasses voxelization methods, pillar encoding, and two detection head approaches (anchor-based and center-based) to develop a state-of-the-art pedestrian detection system aligned with recent advancements in 3D object detection research.

3.3. Benchmarking protocol

A benchmarking protocol has been designed to adapt 3D object detection tasks for video surveillance. It evaluates models across various distance ranges defined in Tab. 2, with a focus on approximately 20 meters, relevant to video surveillance. The evaluation metric is the pedestrian mean average precision (mAP), adapted for 3D object detection,

using nuScenes' methodology [2]. This methodology replaces traditional Intersection over Union (IoU) with 2D center distance thresholding, ensuring accurate evaluation regardless of object size or orientation. The protocol aims to provide a comprehensive evaluation of models in a video surveillance context.

pedestrian mAP				
0-10 (m)	0-20 (m)	0-30 (m)	0-40 (m)	0-50 (m)

Table 2. Metric and evaluation ranges defined by the benchmarking protocol.

Additionally, the protocol considers several critical hyperparameters:

- **Training range**: This defines the maximum range for training annotations, limiting them to a specific distance (e.g., 30 meters).
- **Voxel size**: It determines the size of the voxel used in point cloud processing.
- **Number of sweeps per sample**: This influences the point cloud density by specifying the number of consecutive sweeps in a sample, adding a temporal dimension.
- **Classes used**: It defines whether the model is trained with only pedestrian data or includes all available classes. These hyperparameters play a crucial role in optimizing model performance for video surveillance.

Based on our exploration and experiments with state-of-the-art models using the benchmarking protocol, we have developed an optimized pipeline (Fig. 1) for pedestrian detection in video surveillance scenarios.

4. Experiments

In this section, we present a detailed overview of the diverse experiments conducted to thoroughly evaluate our pedestrian detection methodology. Our experimental exploration encompasses two primary datasets: the nuScenes dataset and the L-CAS dataset, which will be qualitatively analyzed.

SOTA Models Comparison and Tuning. The four state-of-the-art models studied at the methodology were evaluated for pedestrian detection, focusing on the systematic variation of hyperparameters as outlined in the benchmarking protocol. These hyperparameters included voxel size, the number of sweeps, training range, and the number of classes.

The experiments began with the mini nuScenes dataset for efficient hyperparameter grid search, followed by training with the full dataset to achieve optimal results.

Multi-modality. The BEVFusion framework is used to combine camera and LiDAR data in a shared Bird's Eye View (BEV) space. Unlike traditional methods that map one modality to another, BEVFusion preserves the semantic richness of camera data while retaining the geometric structure of LiDAR. This fusion is expected to enhance accuracy, especially for detecting pedestrians, by providing a denser representation compared to LiDAR alone.

The experiment employs a Swin Transformer for image encoding, using pretrained weights from the nuImages dataset. LiDAR encoding is handled by the best-performing model from previous experiments.

Qualitative analysis. Given the lack of sufficient annotations in the L-CAS dataset, a qualitative analysis was primarily employed. This visual inspection of the model's performance offers insights into its indoor detection capabilities. It is important to note that qualitative analysis was also conducted for the nuScenes dataset, while for L-CAS, it remains the primary approach due to missing annotations.

The challenges of adapting between datasets with varying LiDAR devices are recognized. Pre-training models on one dataset and expecting seamless performance on another is found to be impractical due to these disparities. To address this, the best-performing model from nuScenes was chosen and fine-tuned using the L-CAS dataset, acknowledging the domain adaptation complexities.

4.1. Results

SOTA Models Comparison and Tuning. Several valuable insights were gained through model comparison and hyperparameter tuning in Tab. 3:

- **Training Range Influence**: Expanding the training range to 50 meters positively impacted results, indicating that a broader training range enhances model generalization, particularly for metrics relevant to video surveillance.
- **Sweep Number Considerations**: Surprisingly, using 5 sweeps per sample outperformed the recommended 10 sweeps, suggesting that an excessive number of sweeps may not be necessary, especially for close-range detections.
- **Optimal Voxel Size**: Smaller voxel sizes were preferred across all models tested due to improved representation, but they come at the cost of increased inference time.

Training range (m)	Max sweeps	Voxel size	Classes	pedestrian mAP				
				0-10 (m)	0-20 (m)	0-30 (m)	0-40 (m)	0-50 (m)
50	5	[0.08, 0.08, 0.2]	pedestrian	**0.8991**	**0.8763**	**0.8272**	**0.8087**	**0.7811**
30	5	[0.08, 0.08, 0.2]	pedestrian	0.8977	0.8724	0.807	0.7249	0.6547
20	5	[0.08, 0.08, 0.2]	pedestrian	0.8473	0.8137	0.5624	0.4608	0.4088
50	10	[0.08, 0.08, 0.2]	pedestrian	0.8282	0.7949	0.7601	0.7379	0.702
50	5	[0.08, 0.08, 0.2]	pedestrian	**0.8991**	**0.8763**	**0.8272**	**0.8087**	**0.7811**
50	1	[0.08, 0.08, 0.2]	pedestrian	0.8835	0.841	0.779	0.7325	0.6922
50	5	[0.16, 0.16, 0.2]	pedestrian	0.7298	0.7405	0.71	0.6895	0.6587
50	5	[0.1, 0.1, 0.2]	pedestrian	0.868	0.8526	0.8013	0.78	0.7533
50	5	[0.08, 0.08, 0.2]	pedestrian	**0.8991**	**0.8763**	**0.8272**	**0.8087**	**0.7811**
50	5	[0.08, 0.08, 0.2]	all	0.8558	0.822	0.7805	0.7636	0.7366
50	5	[0.08, 0.08, 0.2]	pedestrian	**0.8991**	**0.8763**	**0.8272**	**0.8087**	**0.7811**

Table 3. CenterPoint validation results on the mini nuScenes dataset on different hyperparameter configurations. This table only presents a portion of the grid search that was conducted for the experiments on one of the selected models. The behaviour found for the rest of the models was the same. The optimal hyperparameters found are 50m training range, 5 sweeps, 0.08 voxel size and only using pedestrian class. Blue text indicates the variations of one hyperparameter in the optimal configuration.

- **Single-Class Training**: Training models exclusively with pedestrian data consistently outperformed multi-class training, highlighting the viability of dedicated pedestrian datasets for video surveillance tasks.

- **Model Performance and Representation**: CenterPoint utilizing voxel representation emerged as the highest-performing model, outperforming SECOND and models employing pillar representation. PillarNet's claim of surpassing CenterPoint's results was not supported by the pedestrian benchmark.

The top-performing model was CenterPoint using voxel representation in Tab. 4, achieving an mAP@0-20m score of 0.8763 with optimized hyperparameters. Further training on the complete dataset is planned to explore its full potential.

However, it is important to note that due to limitations in the test set's ground truth, a direct comparison with the state-of-the-art using the benchmark on this test set is challenging. An alternative approach was adopted to compare results from pretrained models, trained on the complete dataset.

The comprehensive training on the best CenterPoint configuration and the complete dataset revealed in Tab. 5 that an expanded training dataset significantly enhances model generalization. This observation emphasizes the importance of dataset size in achieving superior model performance.

Additionally, the optimized model demonstrated noteworthy performance, even surpassing most pretrained weights from state-of-the-art models, underlining the crucial role of the voxel size hyperparameter in model optimization, although achieving the ideal voxel size is influenced by hardware resource constraints.

Multi-modality. Our findings indicate that a smaller window size yielded better performance, indicating an increased representation power and achieving the best results using a window size of 7 pixels.

Comparing the results with the single-modality approach, there is a marginal enhancement within the 10 and 20 meter ranges, visible in Tab. 6. This suggests that multimodality indeed proves advantageous in the context of video surveillance, particularly for closer instances. It is worth noting that instances in closer proximity to the sensor are often more visible in RGB images and consist of a higher point density, likely contributing to improved detection.

However, for greater distances, we observed no substantial improvement, even though the performance remained comparable to that of the single-modality setup. Our hypothesis is that objects in closer ranges are more distinctly captured and aligned between LiDAR and RGB data, facilitating the enhancement in detection performance within these ranges.

In summary, the results of the multimodality experiment, conducted on the mini nuScenes dataset, showed moderate improvements. However, it is worth noting that these improvements were not as significant as those reported in some state-of-the-art multimodal models. This could be due to the limitations of the mini dataset in capturing the full potential of multimodality and the use of a LiDAR encoder

Model	Training range (m)	Max sweeps	Voxel size	Classes	pedestrian mAP			Inference time (ms)
					0-10 (m)	0-20 (m)	0-30 (m)	
SECOND	50	5	[0.08, 0.08, 0.2]	pedestrian	0.8084	0.8134	0.7406	300
CenterPoint	50	5	[0.08, 0.08, 0.2]	pedestrian	**0.8991**	**0.8763**	**0.8272**	261
CenterPoint-P	30	5	[0.1, 0.1, 0.2]	pedestrian	0.8696	0.8125	0.7351	**93**
PillarNet	50	5	[0.08,0.08,0.2]	pedestrian	0.8384	0.8075	0.7443	168

Table 4. Hyperparameter optimization results for every selected model. Configuration with the best validation results from the mini nuScenes dataset are shown for every selected model. Experiments are executed on a Nvidia RTX A2000 12GB.

Model	Training range (m)	Max sweeps	Voxel size	Classes	pedestrian mAP		
					0-10 (m)	0-20 (m)	0-30 (m)
CenterPoint (Ours)	50	5	[0.08, 0.08, 0.2]	pedestrian	**0.9737**	0.9491	0.9161
CenterPoint	-	10	[0.1, 0.1, 0.2]	all	0.94905	0.93841	0.91349
CenterPoint	-	10	[0.08, 0.08, 0.2]	all	0.9636	**0.9525**	**0.9241**
CenterPoint-P	-	10	[0.2, 0.2, 0.8]	all	0.93763	0.91727	0.88813
SECOND	-	10	[0.1, 0.1, 0.2]	all	0.92078	0.91754	0.8921

Table 5. Validation results from the mini nuScenes dataset. Comparison of our optimized CenterPoint model against the available pretrained weights of the state of the art. All the models in this table were trained with the full nuScenes dataset.

Model	pedestrian mAP		
	0-10 (m)	0-20 (m)	0-30 (m)
CenterPoint	0.8991	0.8763	0.8272
BEVFusion	**0.9084**	**0.8835**	**0.8297**

Table 6. Validation results from the mini nuScenes dataset. Comparison between the top-performing single-modality model and the top-performing configuration (window size=7) trained on the multi-modality framework.

optimized for single modality, rather than one tailored for multimodal fusion.

Qualitative analysis. In the qualitative analysis of the nuScenes dataset, the study focused on the outputs of the top-performing CenterPoint model. The analysis revealed that the confidence threshold parameter had a significant impact on the model's performance. When a low confidence threshold of 0.1 was used, there was a noticeable presence of false positive detections. On the other hand, raising the confidence threshold to 0.4 led to a substantial reduction in false positive instances, indicating that increasing the threshold improved detection accuracy. At the same time, the quality of orientation estimations was observed to be suboptimal and imprecise, which highlighted the challenge of accurately predicting orientation in pedestrian detection.

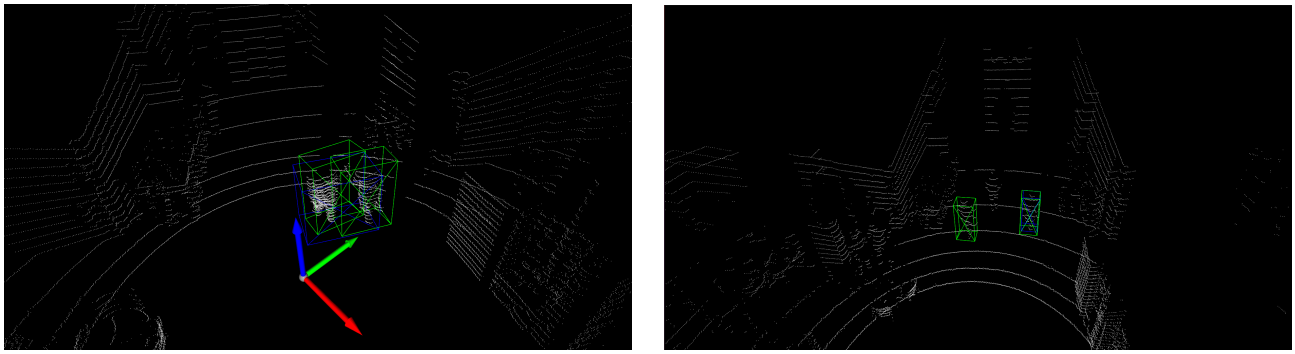
On the other hand, given the constrains presented by the L-CAS dataset, its qualitative analysis employed the top-performing CenterPoint model, fully fine-tuned on the L-CAS dataset. Despite the challenges posed by misannotations in the dataset, several noteworthy observations were made. The model demonstrated the ability to accurately detect pedestrians even in cases where corresponding ground truth annotations were lacking in Fig. 3b. Additionally, in instances where closely positioned pedestrians were annotated as a single group, the model was observed to successfully recognize and estimate these grouped pedestrians as separate entities in Fig. 3a.

In summary, while the L-CAS dataset presented challenges due to misannotations, it also showed promise for addressing novel challenges in indoor LiDAR data analysis, emphasizing its value for future enhancements and deep learning methodologies.

4.2. Ablation Study

In the ablation study, we investigated the effects of annotation simplification on our pedestrian detection model's performance. This study involved three key modifications to the nuScenes annotations, aiming to understand the importance of specific attributes for optimal 3D detection outcomes.

- **No Orientation**: We removed orientation information to assess its importance in detection accuracy.
- **No Velocity**: Velocity data was excluded to investigate its role in accurate detection.



(a) Close pedestrians are sometimes annotated as a singular group.

(b) Missing annotations sometimes are accurately detected.

Figure 3. Qualitative results in the L-CAS dataset. Blue boxes indicate ground truth, and green indicates our detection. Open3D [19] is used for visualization.

- **Cylindrical Detections**: Pedestrian boxes were simplified into cylinders to reduce the number of box parameters. Given the fact that width and length from pedestrians are usually very similar, such are simplified into a single diameter.

Results indicate that predicting orientation and velocity for pedestrians does not significantly improve pedestrian detection performance, regardless of whether the model is trained with all classes or exclusively pedestrian data. This conclusion is based on experiments using the CenterPoint model with specific hyperparameters.

Estimating orientation and velocity for pedestrians in the mini dataset was challenging, as shown in Tab. 7. These difficulties explain why integrating velocity and orientation information did not lead to substantial improvements in pedestrian detection performance. The differences observed among different combinations of these attributes were minimal and may be attributed to random variations. The study relied on two main metrics, the Mean Absolute Orientation Error (mAOE) and the Mean Absolute Velocity Error (mAVE), to assess the accuracy and precision of orientation and velocity predictions within the models.

Additionally, the study noted that using a cylindrical bounding volume did not lead to notable changes in performance. This observation suggests that a simplified bounding volume could be considered for future dataset annotations without compromising detection accuracy.

Orientation	Velocity	ped. mAP	ped. mAOE	ped. mAVE
False	False	0.7904	-	-
False	True	0.7896	-	0.9187
True	False	0.7889	1.638	-
True	True	0.7719	1.518	0.9102

Table 7. Validation result from the mini nuScenes dataset at the range of 20 meters. Showcasing mAP, mAOE and mAVE for pedestrian depending on the annotation simplification. The higher mAP the better, while the lower mAOE and mAVE the better.

5. Conclusions

This work has showcased the effectiveness of voxel-based models for 3D object detection, particularly in the context of pedestrian detection within video surveillance. The use of sparse convolutions, aligned with state-of-the-art research, proved to be a successful approach. Notably, we have identified optimized hyperparameter configurations tailored specifically for pedestrian detection, breaking away from conventional setups for autonomous driving.

Our exploration with the nuScenes dataset revealed that pedestrian detection can be achieved effectively without the need for complex annotations, suggesting the potential for creating a novel dataset designed specifically for video surveillance applications. Additionally, our study on multi-modality demonstrated minor performance enhancements in critical surveillance ranges, calling for further investigation in this complex field.

AI ethics, particularly in surveillance, are crucial. RGB systems may breach privacy, while LiDAR provides spatial data without compromising anonymity. LiDAR's accurate distance data aids in crowd control, intrusion detection, and scene understanding, improving security and decision-making. Future research should involve enhancing annotations in datasets like L-CAS for better applicability in video surveillance deep learning models. Creating a custom dataset for surveillance needs is another promising direction, despite the high cost of LiDAR sensors.

Acknowledgements. This work has been partially supported by the Spanish project PID2022-136436NB-I00 and by ICREA under the ICREA Academia programme.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022. 3

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 3, 4, 5

[3] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 4

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 4

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3, 4

[6] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, pages 409–418. PMLR, 2021. 4

[7] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 1, 2

[8] 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. *arXiv preprint arXiv:2205.13790*, 2022. 3

[9] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[10] Guangsheng Shi, Ruifeng Li, and Chao Ma. Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 35–52. Springer, 2022. 2, 4

[11] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 2

[12] Weijing Shi and Raj Rajkumar. 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*, pages 1711–1719, 2020. 2

[13] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019. 3

[14] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 4

[15] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 3, 4

[16] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2, 4

[17] Zhi Yan, Tom Duckett, and Nicola Bellotto. Online learning for human classification in 3d lidar-based tracking. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 864–871. IEEE, 2017. 4

[18] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 1, 2, 4

[19] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 8

[20] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 2