MissVoxelNet: 3D Object Detection for Autonomous Vehicle in Snow Conditions

Anh Do The
Department of Information Communication Convergence
Soongsil University
Seoul, Korea Republic of.
theanhdo20@soongsil.ac.kr

Myungsik Yoo
School of Electronic Engineering
Soongsil University
Seoul, Korea Republic of.
myoo@ssu.ac.kr

***Abstract*—LiDAR (Light Detection and Ranging) sensors are widely used in self-driving cars with awareness of the surrounding environment. However, the LiDAR sensor is sensitive to harsh weather conditions that cause the collected data to be distorted. These types of weather reduce the safety of self-driving cars. The harsh weather conditions also cause missing points problems on the point clouds, and it causes the performance of 3D object detection to reduce. Therefore, we propose a new method using probability estimation, which includes a Deep Mixture of Factor Analyzers (DMFA) and a Miss-Convolution layer, to recover missing points caused by snow. The proposed work outperforms models which perform well in normal conditions. In summary, snow often causes detection errors for 3D modern detectors. By recovering missing points in the point cloud, we significantly make the performance of the 3D detector better in snowy weather conditions.**

***Index Terms*—Autonomous vehicles, LiDAR, 3D object detection, snowy weather conditions**

I. Introduction

To deliver correct environmental awareness, autonomous vehicle systems rely mainly on precise sensor data, employing multi-sensor setups and pricey sensors such as LiDAR. A 3D scanner known as LiDAR, which stands for Light Detection and Ranging, uses light in laser pulses to detect range, thanks to the rapid development of 3D sensing technology. One of the most significant challenges in developing driverless vehicles and driver assistance systems is how poorly they operate in snowy conditions. Snowy weather conditions also harm LiDAR sensors. As a result, inaccurate sensor data might lead to erroneous decisions and car accidents. Therefore, our main target is to improve the accuracy of the 3D object detection model in snowy weather conditions.

LiDAR sensors are known to be sensitive to adverse weather conditions such as snow due to reduced signal-to-noise ratio and signal-to-background ratio as well as large backscattered power from random droplets. This led to weather-dependent changes in reflectivity and increased range uncertainty. In some cases, this can cause false detection if the signal is reduced below the noise level. Further, the background strength increases, especially close to the sensor, simply because a larger fraction of the backscattered laser power from random droplets is available. Harsh weather causes missing points for the point cloud, which leads to a decrease in the performance of the 3D object detection model. Therefore, our proposed work apply a Deep Mixture of Factor Analyzers (DMFA) and a Miss-Convolution layer to recover missing points. As a result, our proposed model outperforms both of model [7] and model [10].

II. Methodology

A. Overall Architecture

The proposed model has almost parts same as the teacher model in SE-SSD [4]. We put more a Deep Mixture of Factor Analyzers (DMFA) network [3] and a Miss-Convolution layer. The overall proposed architecture is shown in Fig. 1. The model starts with a VoxelNet [6] network, divides the point cloud into box cells, and then uses Voxel Feature Encoding (VFE) to encode into sparse voxel features. Sparse convolution is used to learn information about the z-axis and convert the sparse 3D voxel into a 2D BEV image. BEV means bird's eye view. Sparse convolution network (SpconvNet) has four blocks ($\{2, 2, 3, 3\}$ submanifold sparse convolution [8] layers) with a sparse convolution layer [9] at the end. Then, We apply the DMFA network [3], and Miss-Convolution layer to recover missing points on point cloud. Next, we concatenate the sparse 3D feature along z into a 2D dense feature for feature extraction with the Spatial-Semantic Feature Aggregation (SSFA) module and Attentional Fusion (AF) module. In this work, we use a single shot detector (SSD)-like [5] architecture to do the object detection task. Finally, three 1 × 1 convolutions are applied for label classification, location regression, and direction classification.

B. Miss-Convolution

A missing voxel is indicated by $x = (x_o, x_m) \in \mathbb{R}^n$, where $x_o \in \mathbb{R}^d$ express for voxels with known values, while $x_m \in \mathbb{R}^{n-d}$ denotes loss voxels. The set of indices (voxels) with loss values in sample x is indicated $\mathcal{J} \subset \{1, \ldots, n\}$. While conditional density $p_{x_m|x_o}$ is defined on $(n-d)$-th space, we expand it to the whole $\mathbb{R}^n$ space.

$$P_{x_m|x_o}(t) = \begin{cases} p_{x_m|x_o}(t_{\mathcal{J}'}), & \text{if } t_{\mathcal{J}'} = x_o. \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where $t_{\mathcal{J}'}$ denotes the limitation of $t \in \mathbb{R}^n$ to the observed voxels $\mathcal{J}' = \{1, \ldots, n\} \backslash \mathcal{J}$.

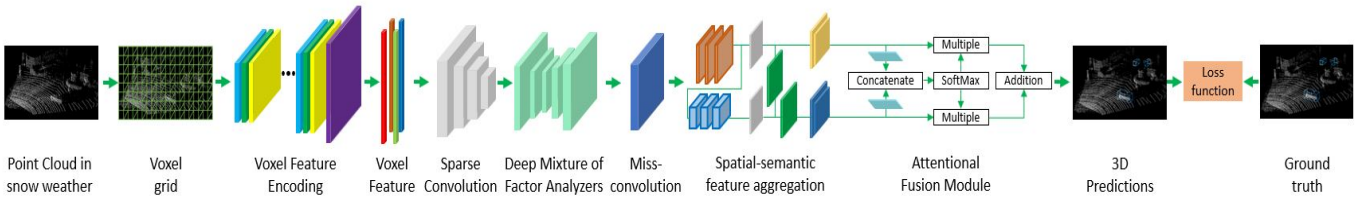

Fig. 1. The architecture of our proposed MissVoxelNet model. The input point cloud will be encoded into features through VoxelNet and Sparse Convolution. Then the Deep Mixture of Factor Analyzers network and miss-convolution layer has the function to recover the lost data. The predictions are made based on the performance of Spatial-Semantic Feature Aggregation and Attentional Fusion modules.

The Factor Analyzer (FA) is a component of the DMFA, and it has a Gaussian distribution with a low-dimensional covariance matrix. A single FA calculated in $\mathbb{R}^n$ is defined by the mean vector $\mu \in \mathbb{R}^n$, and the covariance matrix $\Sigma = AA^T + D$, where $A_{n\times l}$ is the rank factor loading matrix low consists of l vector $a_1, \ldots, a_l \in \mathbb{R}^n$, such that $l \ll n$, and $D = D_{n\times n} = diag(d)$ is a diagonal matrix representing the noise regardless of $d \in \mathbb{R}^n$. FA is formalized as a random vector with the following properties:

$$Z = \mu + \sqrt{d} \odot X + \sum_{j=1}^{l} Y_j \cdot a_j. \tag{2}$$

where $X \sim N(0, I)$, $Y_j \sim N(0, 1)$ are independent, $\sqrt{d}$ denotes element-wise square root of vector d, and $a \odot b$ stands for element-wise multiplication of vectors a and b. We consider a random vector Z with an MFA distribution P_Z representing a missing data point $x = (x_o, x_m)$. Let M be a linear convolution computation, which creates a random vector MZ. The random vector MZ then has an FA distribution with mean and variance calculated as follows:

$$\begin{aligned} &\mathbb{E}[MZ] = M\mu \\ &\mathbb{V}[MZ] = diag(Md) + \sum_{j=1}^{l} (Ma_j) \cdot (Ma_j)^T. \end{aligned} \tag{3}$$

The activation function is then applied to all coordinates of the feature map created by the MZ. The 1-dimensional Gaussian density is $P = \sum_{i=1}^{k} p_i N(m_i, \sigma_i^2)$. When ReLU is applied to a random variable with density P, the predicted value is:

$$\begin{aligned} &\mathbb{E}[ReLU(P)] \\ &= \frac{1}{2} \sum_{i=1}^{k} p_i \Big(m_i + \frac{\sigma_i}{2\sqrt{2\pi}} \exp(-\frac{m_i^2}{2\sigma_i^2}) + m_i \cdot erf(\frac{m_i}{\sigma_i \sqrt{2}}) \Big). \end{aligned} \tag{4}$$

where $erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2) dt$ is the error function. We create a neural network that uses an incomplete x point cloud to return the conditional DMFA parameters. The log-likelihood loss is used to train an inpainting network.

C. Loss Function

Bounding box regression loss, label classification loss, and direction classification loss are the three loss functions in our model. This model is trained using typical loss functions in object detection. For the bounding box position and angle regression task of the teacher model, we utilize the Smooth-L1 loss function L_{reg}:

$$\begin{aligned} L_{reg} &= SmoothL1(\delta_b) \\ \delta_b &= \begin{cases} |b_p - b_{gt}|, & \text{if } b \in \{x, y, z, w, l, h\} \\ |\sin(b_p - b_{gt})|, & \text{if } b \in \{r\} \end{cases} \end{aligned} \tag{5}$$

where $\{x, y, z\}$, $\{w, l, h\}$, and r denote the center position, sizes, and orientation of a bounding box, respectively, subscript p means prediction, subscript gt means ground truth. The label classification task is given by a focal loss L_{cls}:

$$\begin{aligned} L_{cls} &= -\alpha(1 - \delta_l)^\gamma \log(\delta_l) \\ \delta_l &= |\sigma(l_p) - \sigma(l_{gt})| \end{aligned} \tag{6}$$

where α and γ are the parameters of the focal loss. The sigmoid classification scores of prediction and ground truth are $\sigma(l_p)$, $\sigma(l_{gt})$, respectively.

The direction classification loss L_{dir} is calculated using the softmax function. To generate a direction classification target, we utilize the following method: if the rotation around the z-axis of the ground truth is less than 0, the value is negative; otherwise, the value is positive.

III. Experiment and Result

A. Experimental Setup

1) Dataset: The production of new synthetic datasets in snow is demonstrated in this section. We used the LISA simulator [2] to build new synthetic datasets in snow conditions based on the KITTI point cloud data [1]. We got new datasets called Snow-KITTI, which are synthetic point cloud datasets for snow condition. The weather type is divided into three categories: light, medium, and heavy, as illustrated in Tab. II. The modification of the point cloud in snow circumstances with intensity levels of light, medium, and heavy is shown in Fig. 2. In snow, there are some noises near the LiDAR sensor's position and missing point objects in point cloud. Following the typical methodology, each of our training datasets comprises 3,712 training samples and 3,769 validation samples.

TABLE I
AVERAGE PRECISION OF 3D OBJECT DETECTION COMPARISONS SPLIT ACROSS THREE SNOW INTENSITY LEVELS AND THREE DIFFICULTY LEVELS [%]. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	Light			Medium			Heavy		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Pointpillar [7]	43.96	30.27	28.53	38.66	26.69	24.20	33.19	22.37	20.89
SECOND [11]	69.43	50.08	46.72	62.93	44.24	40.56	57.05	35.28	31.95
def PV-RCNN [10]	73.71	55.50	50.53	68.85	48.64	44.32	61.29	41.77	38.67
MissVoxelNet (our)	**75.82**	**56.04**	**50.92**	**70.34**	**50.61**	**44.88**	**63.66**	**44.63**	**39.61**

TABLE II
THE RATES OF SNOWFALL BASED ON LEVELS OF SIMULATED SNOWY WEATHER INTENSITY

Levels of intensity	Light	Medium	Heavy
The rate of snowfall (mm/hr)	0 - 0.8	0.8 - 0.9	> 0.9

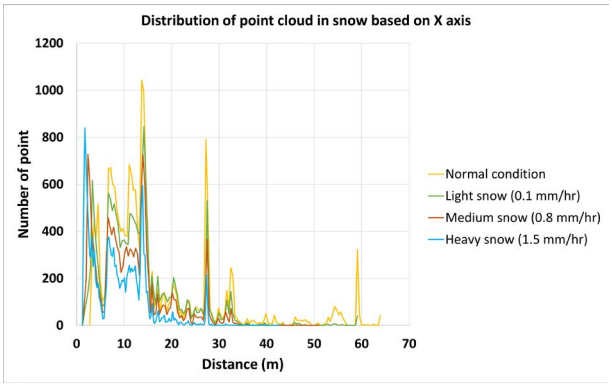

Fig. 2. The histogram shows the point distribution of point cloud in snowy weather conditions based on X-axis.

2) Evaluation Metrics: We evaluate the performance of 3D object detection models according to three levels of snow intensities: light, medium, heavy. To measure the performance of the 3D object detection task, average precision (AP) is computed across 40 recall positions values between 0 and 1. Our trials are primarily focused on the most often used car category, and the average precision is measured using an Intersection over Union (IoU) threshold of 0.7. The benchmark also contains three difficulty levels in the evaluation: easy, moderate, and hard, which are dependent on object size, occlusion, and truncation levels, with moderate average precision being the official ranking statistic for 3D detection.

3) Training Details: We implemented our model with PyTorch (1.8) and trained on a RTX 3060 GPU. During the training, we choose Adam Optimizer with learning rate $1e^{-3}$ with batch size 6 and total epoch 150.

B. Results

The MissVoxelNet model's results in 3D object detection were compared with Pointpillar [7], SECOND [11] and def PV-RCNN [10] model, which performed well in normal conditions. For the comparison, we use the same input from SnowKITTI for all models. The comparison results are shown in Tab. I. The performance of Pointpillar [7], SECOND [11] and def PV-RCNN [10] model in snowy weather conditions were reduced. The leading cause is losing data. By recovering the lost points based on combining the probability estimation method with the existing architecture of 3D object detection, we have improved the model's performance. It can be seen that the performance of the proposed model is far superior and achieves the best results in all snow intensity levels and object difficulty levels.

IV. CONCLUSION

Using a new network called MissVoxelNet, we offer a novel method for 3D object detection in this research. By incorporating a DMFA network [3] and a Miss-Convolution layer into an existing model, missing points on the point cloud can be recovered. As a result, when compared to other methodologies, our model outperforms the competition.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Goverment of South Korea (MSIT)(NRF-2021R1A2B5B01002559)

REFERENCES

[1] Geiger, Andreas; LENZ, Philip; URTASUN, Raquel. Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012. p. 3354-3361.
[2] Kilic, Velat, et al. Lidar Light Scattering Augmentation (LISA): Physics-based Simulation of Adverse Weather Conditions for 3D Object Detection. arXiv preprint arXiv:2107.07004, 2021.
[3] Przewięźlikowski, Marcin; ŚMIEJA, Marek; STRUSKI, Łukasz. Estimating conditional density of missing values using deep gaussian mixture model. In: International Conference on Neural Information Processing. Springer, Cham, 2020. p. 220-231.
[4] Zheng, Wu, et al. SE-SSD: Self-ensembling single-stage object detector from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. p. 14494-14503.

[5] Liu, Wei, et al. Ssd: Single shot multibox detector. In: European conference on computer vision. Springer, Cham, 2016. p. 21-37.
[6] Zhou, Yin; TUZEL, Oncel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 4490-4499.
[7] Lang, Alex H., et al. Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p. 12697-12705.
[8] Graham, Benjamin; ENGELCKE, Martin; VAN DER MAATEN, Laurens. 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 9224-9232.
[9] Liu, Baoyuan, et al. Sparse convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 806-814.
[10] Bhattacharyya, Prarthana; CZARNECKI, Krzysztof. Deformable PV-RCNN: Improving 3D object detection with learned deformations. arXiv preprint arXiv:2008.08766, 2020.
[11] Yan, Yan; MAO, Yuxing; LI, Bo. Second: Sparsely embedded convolutional detection. Sensors, 2018, 18.10: 3337.