IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 9 December 2023, accepted 4 January 2024, date of publication 9 January 2024, date of current version 22 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3351868

RESEARCH ARTICLE

Enhancing Indoor Robot Pedestrian Detection Using Improved PIXOR Backbone and Gaussian Heatmap Regression in 3D LiDAR Point Clouds

DUY ANH NGUYEN[1], KHANG NGUYEN HOANG[2], NGUYEN TRUNG NGUYEN[2], DUY ANH NGUYEN[1], AND HOANG NGOC TRAN[2]

[1]Department of Mechatronic Engineering, Faculty of Mechanical Engineering, Ho Chi Minh City University of Technology (HCMUT), Vietnam National University Ho Chi Minh City, Vietnam

[2]Department of Software Engineering, FPT University, Can Tho 94000, Vietnam

Corresponding author: Hoang Ngoc Tran (hoang2531992@gmail.com)

ABSTRACT Accurate and robust pedestrian detection is fundamental for indoor robotic systems to navigate safely and seamlessly alongside humans in spatially constrained, unpredictable indoor environments. This paper presents a novel method, IRBGHR-PIXOR, a detection framework specifically engineered for pedestrian perception in indoor mobile robots. This novel approach employs an enhanced adaptation of the cutting-edge PIXOR model, integrating two pivotal augmentations: a remodeled convolutional backbone leveraging Inverted Residual Blocks (IRB) in unison with Gaussian Heatmap Regression (GHR), as well as a Modified Focal Loss (MFL) function to tackle data imbalance issues. The IRB component notably bolsters the network's aptitude for processing intricate spatial representations generated from sparse 3D LiDAR scans. Meanwhile, integrating GHR further elevates accuracy by enabling precise localization of pedestrian subjects. This is achieved by modeling the probability distribution and predicting the central location of individuals in the point cloud data. Extensively evaluated on the large-scale JRDB dataset comprising intense scans from 16-beam Velodyne LiDAR sensors, IRBGHR-PIXOR accomplishes exceptional results, attaining 97.17% Average Precision (AP) at the 0.5 IOU threshold. Notably, this level of accuracy is achieved without significantly increasing model complexity. By enhancing algorithms to overcome challenges in confined indoor environments, this research paves the way for safe and effective deployment of autonomous technologies once encumbered by perceptual limitations in human-centered spaces. Nonetheless, evaluating performance in diverse edge cases and integration with complementary sensory cues promise continued progress. The developments contribute towards the vital capacity of reliable dynamic perception for next-generation robotic systems coexisting in human-centric environments.

INDEX TERMS Robot navigation, pedestrian detection, pedestrian tracking, PIXOR, Gaussian heatmap, point cloud, deep learning.

I. INTRODUCTION

In recent years, the deployment of mobile robots in human-occupied indoor spaces has gained significant interest, with diverse applications spanning from automated inventory and surveillance operations in warehouses, to service robotics in retail environments, and assistive technologies for the elderly in residential care [1]. This widespread adoption hinges on the integration of highly precise and dependable pedestrian detection systems into these robots. However, pedestrian detection for indoor mobile robots poses numerous distinct challenges rarely encountered in outdoor navigation. This arises largely due to confined spaces exacerbating the unpredictability of pedestrian movements, coupled with complications from poor lighting, sensor noise interference, infrastructure obstruction, and dense crowds. Consequently, achieving near-flawless pedestrian tracking is crucial, demanding a level of accuracy and reliability that

The associate editor coordinating the review of this manuscript and approving it for publication was Yangmin Li.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

current detection methods struggle to provide. Therefore, specialized solutions are imperative.

As technology has advanced, so too have the methods for pedestrian detection. Traditional camera-based systems, despite undergoing significant advancements [2], continue to confront challenges, particularly with changing lighting conditions and the inherent complexity of depth perception. In contrast, LiDAR technology has brought forth a new dimension in detection capabilities by providing intricate 3D point cloud data that offers a more detailed understanding of the object space [3], [4].

With the introduction of deep learning, pioneering 3D object detection techniques that utilize this point cloud data have emerged, with VoxelNet [5], [6] and PointPillars [7] at the forefront. Among these, PIXOR [8] distinguishes itself as a noteworthy single-stage detector that has shown remarkable proficiency with raw point cloud data, though it encounters some challenges when applied to the complex JRDB dataset [1]. These deep learning approaches, especially those that integrate CNNs [9], have transformed 2D image processing and are now making significant strides in the realm of 3D LiDAR data interpretation [10], [11]. VoxelNet, for instance, has been instrumental in converting point clouds into structured voxel grids, effectively balancing precision with computational efficiency [5]. Nonetheless, there remain challenges in customizing models such as PIXOR—designed to skip preprocessing steps—for use in indoor environments where unique challenges such as spatial limitations, varied lighting conditions, intricate human movement patterns, and sensor noise can complicate detection [12]. The unveiling of the JRDB dataset [13], [14], which captures the complexity of indoor environments, underscores the pressing need to refine pedestrian detection algorithms for such challenging settings.

PIXOR represents a cutting-edge 3D object detection system that processes LiDAR point clouds directly, utilizing a bird's eye view for efficient representation combined with a CNN optimized for precise, density-oriented box predictions [8]. This approach simplifies the 3D point cloud data into a 2D grid, preserving essential height information while reducing the complexity typically associated with 3D voxel grids. Through this technique, PIXOR's architecture extracts multi-scale features, and its head component accurately predicts object classes and geometrical box shapes. The system's innovative features, including its anchor-free approach and the omission of separate region proposal mechanisms, have propelled it to the forefront of the field, evidenced by its stellar performance in benchmarks such as KITTI, where it operates in real-time at rates surpassing 28 FPS [15].

The JRDB dataset provides a unique egocentric view, documenting the experiences of a social robot named JackRabbot as it navigates the diverse indoor and outdoor spaces of a university campus. This dataset is replete with a wide range of human activities and introduces complex challenges due to the variability in scene settings, population densities, and the presence of occlusions. Its comprehensive annotations and extensive data serve as an invaluable resource for developing a deeper understanding of human-centric robotic perception.

Faced with the rich and intricate details of the JRDB dataset, which reflect the multifaceted nature of human activities in indoor environments, even a robust system like PIXOR encounters new challenges. This dataset presents a demanding set of scenarios that necessitate creative and innovative solutions. Our research is directly aimed at addressing these challenges. We endeavor to refine and adapt PIXOR to more effectively handle the specificity and unpredictability of indoor pedestrian detection, ensuring that it can navigate and interpret the complex dynamics of human activities within such environments. Our goal is to tailor PIXOR's capabilities to meet these demands, thereby advancing the field of pedestrian detection and contributing to the safety and efficiency of autonomous systems operating alongside humans.

In our research, we augment the PIXOR model to address the specific challenges of indoor pedestrian detection. Fig. 1 shows the proposed method's framework and the structure of IRBGHR-PIXOR. Below is a summary of our contributions:

1) We develop an enhanced PIXOR backbone tailored for indoor LiDAR datasets, achieved through the combination of the Inverted Residual Block technique and the application of Gaussian heatmap regression [16] within the object detection head (IRBGHR-PIXOR). These improvements optimize the model, significantly boosting its accuracy in detecting objects of varying sizes.
2) A modified focal loss [17] function has been incorporated to handle class imbalances, emphasizing the significance of precise pedestrian localization.
3) In the experimental results, IRBGHR-PIXOR can achieve higher accuracy on the JRDB validation set compared to previous single-stage detectors. Overall, these advancements pave the way for a robust solution to indoor pedestrian detection challenges.

The remainder of this paper is organized as follows. Section II reviews related work on 3D object detection using LiDAR point clouds and other modalities. Section III describes the proposed methods in detail, including the enhanced PIXOR architecture, Gaussian heatmap regression, and modified focal loss. Section IV presents extensive experiments on the JRDB dataset, analyzing the results and benchmarking the performance of IRBGHR-PIXOR against state-of-the-art approaches. Finally, Section V concludes the paper and discusses directions for future work.

II. RELATED WORK

A. OBJECT DETECTION WITH CNN

In computer vision, the detection of objects has seen impressive advancements, particularly due to the emergence of Convolutional Neural Networks (CNN). Initially lauded for their image classification capabilities [18], CNNs have been skillfully modified for object detection tasks. Such

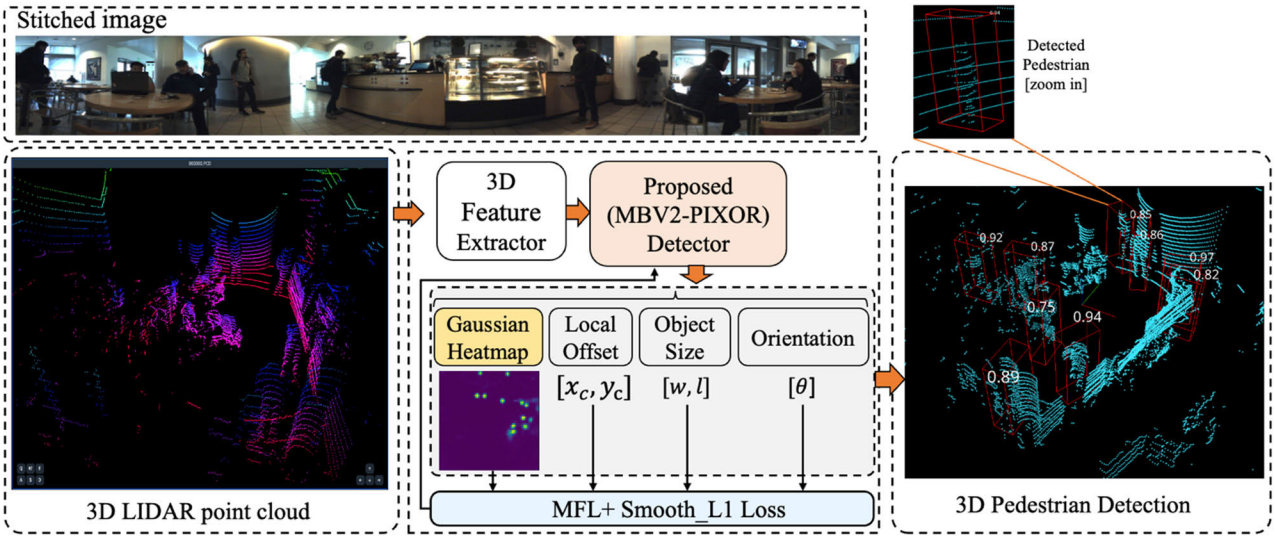

FIGURE 1. Overview of the proposed 3D pedestrian detection framework utilizing LiDAR point cloud data.

modifications typically include the analysis of specific image segments for object presence, a practice showcased by Overfeat [19], which employs a CNN to traverse different areas and scales to determine object boundaries. The innovation of class-independent object proposals [20] has further honed this process, giving rise to key developments such as Region-CNN (RCNN) [21] and faster subsequent versions [22]. Notably, Faster-RCNN [23] has been pivotal in integrating CNNs for both extracting features and formulating object proposals, enhancing the efficiency of the detection process and setting impressive new standards [24]. Yet, the complexity inherent in these dual-stage methods sometimes restricts their deployment in real-time situations. Concurrently, 2D object detection has been substantially improved by CNN methodologies, and there has been a parallel surge in 3D detection technologies, especially with the application of LiDAR systems for pedestrian detection, signifying a vital transition from 2D to the complex analysis and understanding of 3D spatial data.

B. 3D DETECTION OF PEDESTRIANS WITH LiDAR

LiDAR sensors, a critical component in autonomous navigation systems, capture three-dimensional structural data of environments, producing sparse 3D point clouds that detail geometric shapes with high precision [9], [25], [26]. Initially, LiDAR detectors employed manually engineered features and systematically scanned areas using sliding window techniques to detect objects [27]. However, the integration of deep learning has led to a significant leap in the performance of 3D object detection systems. Deep learning frameworks such as PointNet [28] and PointNet++ [29] have been at the forefront, processing unstructured point clouds for classification and segmentation by learning the spatial distribution of points through layers of perceptrons and pooling operations.

Further, voxel-based architectures like VoxelNet [5] and SECOND [30] have advanced the field by converting point clouds into structured voxel grids, which are better suited for convolutional operations, thus facilitating more effective and efficient 3D object detection. Innovations like PointPillars [7] and PointRCNN [10] have introduced methods to condense point clouds into dense representations, optimizing them for 2D convolutional network processing. The development of single-stage detection models such as PIXOR [8] and Point-GNN [31] has streamlined the detection workflow by omitting the region proposal phase, enhancing speed without compromising accuracy. Advanced methodologies, such as PV-RCNN [32], [33], blend the robust features of voxel-based CNNs with the flexibility of PointNet architectures, while PointPainting [34] augments point cloud data with semantic information from other sensors, thereby overcoming the granularity limitations inherent in voxel-based techniques. These deep learning advancements have not only improved the precision of LiDAR-based 3D detectors but have also enabled their application in real-time scenarios. Nonetheless, interpreting human activities within these detailed 3D environments remains a complex challenge that continues to drive research in the field [35].

C. 3D DETECTION OF PEDESTRIANS WITH CAMERA

Camera systems, in contrast to LiDAR, capture the world in two dimensions through RGB imagery, offering rich visual data such as color, texture, and semantic details that complement the depth information provided by LiDAR [36]. The challenge of deriving three-dimensional data from these inherently flat images has been a significant hurdle. Early techniques in monocular 3D detection, like 3DOP [36], utilized geometric constraints to infer depth information. The rise of deep learning has seen CNN-based models like Mono3D [37], [38], [39] and Deep3DBox [40] employ

sophisticated algorithms to extrapolate the third dimension from 2D detections.

Stereo camera setups and depth sensors offer a direct depth perception that aids in the accurate localization of objects in 3D space, with notable approaches such as FQNet [41] and DSGN [42] exploiting this data. Additionally, self-supervised learning methods like MonoDIS [43] forgo the need for exhaustive 3D annotations by learning depth cues from the image data itself. However, these camera-based systems still fall short of the accuracy levels achieved by LiDAR, particularly in estimating depth in monocular setups. The synergy of fusing camera and LiDAR data is an area of intense research, seeking to merge the strengths of both modalities for superior 3D detection capabilities [44]. Despite the strides made in this integration, the passive nature of cameras means they still cannot match the depth-sensing fidelity of active systems like LiDAR and radar on their own [45]. The pursuit of refining passive camera-based detection to match the efficacy of active sensing technologies remains a dynamic and evolving field within computer vision.

III. PROPOSED METHODS

Our developed system employs LiDAR data for precise 3D detection of pedestrians in indoor settings, focusing on accurate localization and heading angle prediction. It utilizes a more efficient 2D representation of LiDAR point clouds for real-time processing. Upcoming sections will discuss improvements to the system's architecture and the implementation of an optimized loss function to boost accuracy.

A. DATA REPRESENTATION

The paper employs the Bird's Eye View (BEV) as its chosen input representation. BEV serves as a way to represent LiDAR point clouds. The decision to favor BEV over a 3D voxel grid is driven by the fact that conventional neural networks operate under the assumption that input data is grid-based. The use of a 3D voxel grid introduces unnecessary computational overhead, especially given the sparse nature of LiDAR point cloud data. BEV representation reduces the input's dimensionality from three to two while preserving height information, allowing the application of efficient 2D convolutions. This approach simplifies object detection, maintains metric space, and enables the network to leverage prior knowledge regarding the physical dimensions of objects.

In the case of a 3D pedestrian of the nth instance, we specify a set of 3D ground truth bounding box attributes as $\left(x^{(n)}, y^{(n)}, z^{(n)}, w^{(n)}, l^{(n)}, h^{(n)}, \theta^{(n)}\right)$. Here, $x^{(n)}, y^{(n)}$ and $z^{(n)}$ represent the central position in the LiDAR coordinate system, while $w^{(n)}$, $l^{(n)}$, and $h^{(n)}$ describe the width, length, and height of the bounding box. Furthermore, $\theta^{(n)}$ denotes the yaw rotation around the z-axis, which is oriented perpendicular to the ground. We establish the 3D spatial dimensions $(x_{min}, x_{max}) \times (y_{min}, y_{max}) \times (z_{min}, z_{max})$ for the area of interest where we aim to detect objects.

For every pedestrian's center in the form of a 2D coordinate in the BEV image system, we obtain the keypoints as follows:

$$p = \left(p_{x^{(n)}}, p_{y^{(n)}}\right) = \left(\frac{x^{(n)} - x_{min}}{r_x . R_{down}}, \frac{y^{(n)} - y_{min}}{r_y . R_{down}}\right) \quad (1)$$

where r_x, r_y are the resolution of per cell, R_{down} is the output stride (downsample ratio of the model). To represent the 2D bounding box in the BEV, we can express it as $\left(p_{x^{(n)}}, p_{y^{(n)}}, \frac{w^{(n)}}{r_x . R_{down}}, \frac{l^{(n)}}{r_y . R_{down}}, \theta^{(n)}\right)$. We will have corresponding heads for various components, including the keypoint Gaussian heatmap head, local offset head, object size head, and orientation head. With the proposed Gaussian heatmap head, the values of each coordinate are explained in the following section.

B. NETWORK ARCHITECTURE

In this study, we have made several significant modifications to the backbone and head of the PIXOR model [8] to improve the performance of our pedestrian detection system. Leveraging the fully convolutional architecture specifically designed for high-density 3D object detection in PIXOR, this network generates pixel-wise predictions in a single stage, where each prediction corresponds to a 3D object estimation. The calculation of these density predictions is performed efficiently and directly, without the need for predefined anchor objects. This design eliminates the requirement for anchor objects to be predefined and avoids the necessity of adjusting the number of proposals transferred from the first stage to the second stage, along with the corresponding Non-Maximum Suppression threshold. This approach has demonstrated effective performance in practical scenarios. Nonetheless, while the original study [8] applied PIXOR for the detection of large-sized vehicles in expansive 3D LiDAR spaces, our research is centered around the detection of pedestrians for indoor robotic applications. This presents a more complex challenge due to the smaller scale of humans within the 2D x-y plane and the intricate interplay of complex backgrounds. To resolve this challenge, we have introduced structural modifications to the PIXOR backbone to enhance its operational efficiency. Furthermore, we have taken into account the inherent correlation between pedestrian objects and their immediate surroundings by employing a 2D Gaussian distribution strategy to represent the ground truth, as opposed to the conventional hard-label approach. In our approach, we treat small objects as keypoints within their relevant context. Additionally, we have modified the corresponding focus loss function to attain a higher level of accuracy in the detection of small-sized objects, surpassing the performance of several other advanced methods. The proposed network architecture is shown in Fig. 2.

1) MODIFIED BACKBONE STRUCTURE

Basing our work on PIXOR's architecture, we have revamped the model's backbone. To begin with, we replaced the

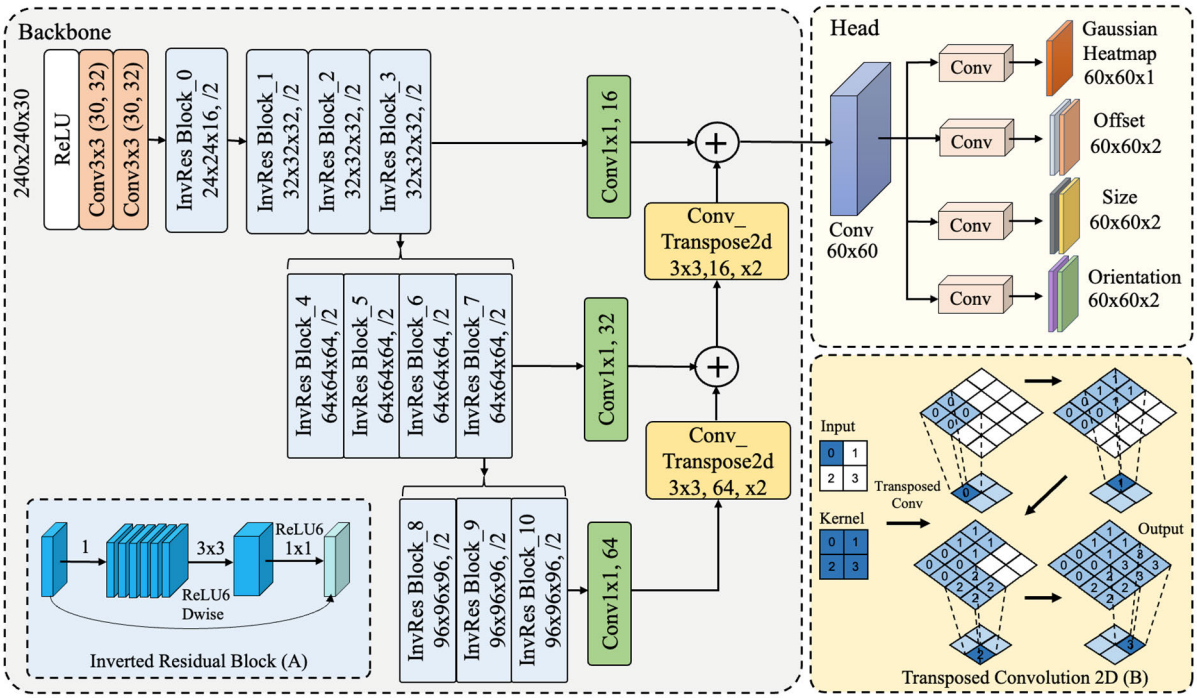

FIGURE 2. The proposed network architecture (IRBGHR-PIXOR) for detecting pedestrians using the BEV of LiDAR 3D point cloud.

Bottleneck blocks with Inverted Residual Blocks. In conventional networks, Bottleneck blocks typically consist of convolutional layers. These convolutional layers serve to extract a comprehensive representation of input features while pooling layers aid in reducing the feature map size, thereby decreasing computational demands and enhancing representation robustness. Nevertheless, these blocks still entail considerable computational complexity, as has been discussed in previous studies [46] and [47].

The utilization of Inverted Residual Blocks reduces the computational workload by a factor of ten. For a visual comparison, please refer to Fig. 3, which illustrates the structural distinctions between the Bottleneck Residual Block and the Inverted Residual Block.

Nevertheless, the challenge becomes more pronounced when dealing with small-sized objects, such as pedestrians in this context. The size of a pedestrian figure is approximately $\mathbf{8} \times \mathbf{12}$ pixels when utilizing a discretization resolution of $\mathbf{0.05}\boldsymbol{m}$. Following a $\mathbf{16}\times$ downsampling process, it occupies less than 1 pixel, rendering the precise identification of pedestrians within this spatial domain quite a daunting task. To surmount this, we devised a strategy to merge information from high-resolution feature maps through skip connections and amalgamated this data with low-resolution feature maps. The up-sampling of the latter was executed using Transpose Conv2d layers with ***stride*** $=$ 2, ultimately resulting in the ultimate sampled representation. Furthermore, we employed Gaussian heatmap regression to address the challenge of dealing with small objects.

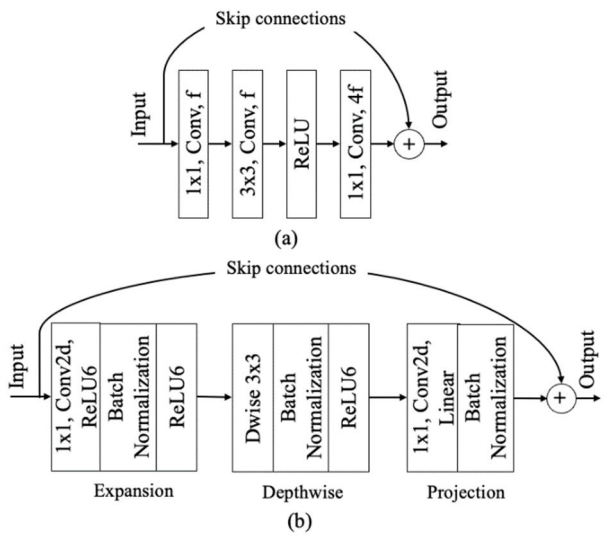

FIGURE 3. The structure of the two main blocks used in PIXOR and the proposed method (IRBGHR-PIXOR): (a) Bottleneck residual block, (b) Inverted residual block.

Our proposed backbone architecture (IRBGHR-PIXOR) is depicted in Fig. 2, encompassing a total of eighteen layers. The first two blocks are constructed as $\boldsymbol{Conv}\mathbf{3} \times \mathbf{3}$ with $\mathbf{32}$ channels each. Subsequently, we have incorporated eleven Inverted Residual Blocks, structured to facilitate feature extraction and data compression in a specific arrangement, with varying quantities $(\mathbf{1}, \mathbf{3}, \mathbf{4}, \mathbf{3})$. The initial convolution within each residual block adopts a stride of $\mathbf{2}$ to facilitate

feature map subsampling. The cumulative outcome is a downsampling factor of **16.** We seamlessly integrate the output of three ***Conv*1 × 1** blocks through skip connections and further refine the result by employing two transposed convolution blocks, thus generating the final feature-sampled map. This ultimately yields the final feature map, exhibiting a **4×** down-sampling factor in relation to the input size (transitioning from **240 × 240 *to* 60 × 60**).

2) HEAD NETWORK STRUCTURE

The head network serves the dual purpose of object recognition and localization. However, since the focus of this study is primarily on detecting pedestrians, the output of the head network for object recognition takes the form of a heatmap representing the likelihood of pedestrians. A sigmoid activation function is employed in this particular head. This predicted heatmap is compared to the ground truth heatmap, which is transformed using a Gaussian kernel during training. The localization head is further divided into three smaller heads, including the local offset head, object size head, and orientation head, which will be discussed in more detail in the following subsection.

a: RECOGNITION HEAD WITH GAUSSIAN HEATMAP

The recognition head employs a Gaussian kernel for each object within the labeled input data, where each object is associated with a keypoint. All the ground truth pixels are transformed into a heatmap denoted as $K_{x,y} \in R^{\left(\frac{x_{max}-x_{min}}{r_x \cdot R_{down}}, \frac{y_{max}-y_{min}}{r_y \cdot R_{down}}\right)}$. The ground truth for the object is defined by a non-standard two-dimensional Gaussian function as follows:

$$\boldsymbol{K_{x,y} = exp\left(-\frac{(x-p_x)^2 + (y-p_y)^2}{2\sigma_p^2}\right)} \tag{2}$$

where $(x - p_x)$ and $(y - p_y)$ are the distances to the center of the object, and σ_p is the standard deviation that adapts to the object's size [48]. In case of an overlap between two Gaussian functions in the conversion process, the one with the greater intensity is chosen. Fig. 4(a) depicts the process of transforming data from a 3D LiDAR point cloud into the ground truth heatmap by Gaussian kernel. Fig. 4(b) exhibits the areas where objects 1 and 2 intersect, and we opt for the object with the higher intensity level. $\hat{K}_{x,y}$ stands for the predicted heatmap value coming from the model using the Gaussian heatmap head. A value of $\hat{K}_{x,y} = 1$ denotes the object's center, while $\hat{K}_{x,y} = 0$ specifies that the pillar is categorized as background.

b: LOCALIZATION HEAD

In the localization head, we have divided it into three separate components to determine the object's offset, size, and orientation. Each object is represented as an oriented bounding box with parameters $\{x_c, y_c, w, l, \theta\}$. Here, (x_c, y_c) indicate the center position of the object while, (w, l) denote the width and length of the object in the x-y plane and,

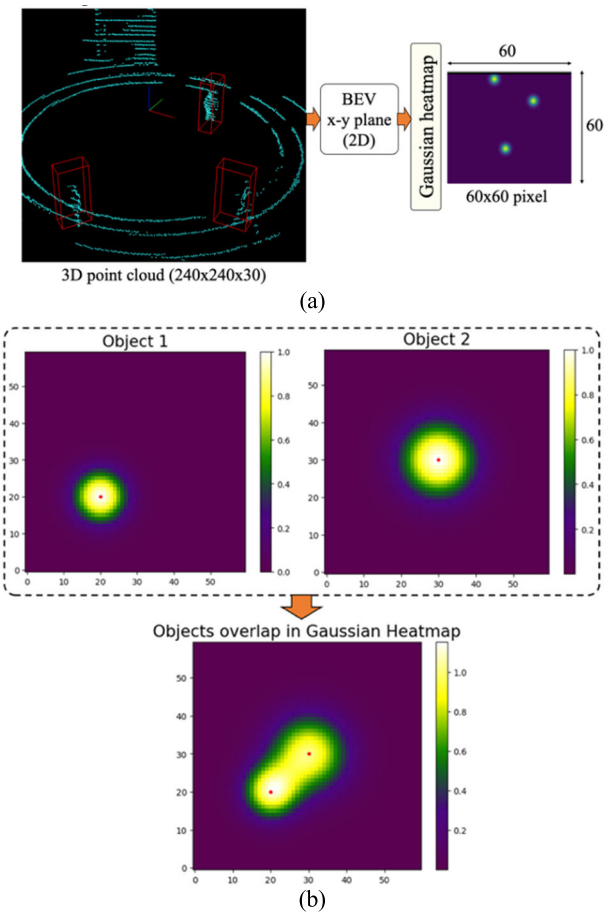

FIGURE 4. Prepare training data: (a) the process of transforming data from a 3D LiDAR point cloud into the ground truth heatmap, (b) the areas where objects 1 and 2 intersect, and the object with the higher intensity level is selected.

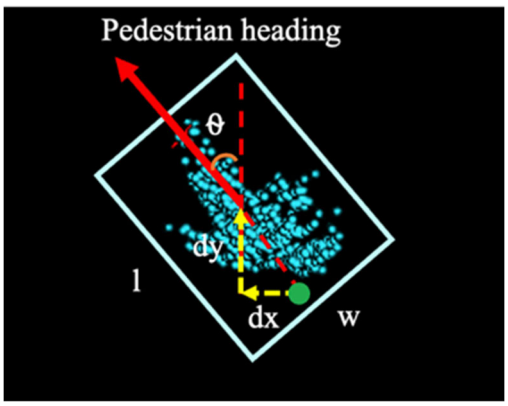

FIGURE 5. The parameterization for a pixel in the geometry output.

θ represents the heading angle of the object ($-\pi$ to π). We do not consider the z-axis as the pedestrian objects we are interested in are constrained to a single plane.

In Fig. 5, the output of the network's regression head is represented as $\{dx, dy, w, l, cos(\theta), sin(\theta)\}$, with each component signifying specific object properties at each pixel location (p_x, p_y). To ensure the heading angle falls within the

desired range, it is decomposed into two correlated values. During inference, θ is decoded using $atan2\left(sin\left(\theta\right), cos\left(\theta\right)\right)$. Importantly, these values pertaining to object position and size are in real-world metric units. By employing the logarithmic function to pre-scale the training dataset, we attain zero mean and unit variance in the learning target $\{log\left(dx\right), log\left(dy\right), log\left(w\right), log\left(l\right), cos\left(\theta\right) sin\left(\theta\right)\}$.

c: LOSS FUNCTION

In this subsection, we will present the proposed loss function used for the pedestrian detector in 3D space. In object recognition, let $\hat{K}_{x,y}$ represent the heatmap value at (x,y) in the predicted heatmap and let $K_{x,y}$ be the heatmap value at (x,y) in the ground truth heatmap. In the previous PIXOR model research [8], a traditional cross-entropy (CE) loss function was employed, as illustrated in (3):

$$\begin{aligned} L_{heat-CE} &= -\frac{1}{N}\sum_{x,y} L_{Cross\ entropy}\left(K_{x,y}, \hat{K}_{x,y}\right) \\ &= -\frac{1}{N}\begin{cases} log\left(\hat{K}_{x,y}\right) & if\, K_{x,y}=1 \\ log\left(1-\hat{K}_{x,y}\right) & else \end{cases} \end{aligned} \tag{3}$$

where N is the total count of objects within the detection range. However, it may not be as effective when there is a significant class imbalance or noisy training data. The Focal Loss (FL) function [49] has been developed with the intention of reducing the loss for well-classified samples and increasing the loss for poorly classified ones as follows (4):

$$\begin{aligned} &L_{heat-FL} \\ &= -\frac{1}{N}\sum_{x,y} L_{Focal\ loss}\left(K_{x,y}, \hat{K}_{x,y}\right) \\ &= -\frac{1}{N}\sum_{x,y}\begin{cases} \alpha\left(1-\hat{K}_{x,y}\right)^{\beta} log\left(\hat{K}_{x,y}\right), & if\, K_{x,y}=1 \\ (1-\alpha)\left(\hat{K}_{x,y}\right)^{\beta} log\left(1-\hat{K}_{x,y}\right), & else \end{cases} \end{aligned} \tag{4}$$

where α and β are the hyperparameters. In the (4), α is a constant. Nevertheless, in this paper, when the ground truth heatmap value deviates from 1, it is required to penalize objects located farther from the center more significantly. Consequently, a modified Focal Loss (MFL) function has been suggested for this purpose, as expressed in the following (5):

$$\begin{aligned} &L_{heat-MFL} \\ &= -\frac{1}{N}\sum_{x,y} L_{ModifiedFocalloss}\left(K_{x,y}, \hat{K}_{x,y}\right) \\ &= -\frac{1}{N}\sum_{x,y}\begin{cases} \left(1-\hat{K}_{x,y}\right)^{\alpha} log\left(\hat{K}_{x,y}\right), & if\, K_{x,y}=1 \\ \left(1-K_{x,y}\right)^{\beta}\left(\hat{K}_{x,y}\right)^{\alpha} log\left(1-\hat{K}_{x,y}\right), & else \end{cases} \end{aligned} \tag{5}$$

The coefficients α and β will be selected through our experimental results. We will evaluate the impact of these loss equations to determine the best equation for accurately detecting pedestrians indoors.

In the localization head, Smooth L1 loss is used as a replacement for the conventional L1 loss to promote stability, and prevent abrupt gradient maxima that may lead to premature convergence or an unstable training process. In the experimental section, we will compare and assess L1, L2, and Smooth L1 losses. The three loss functions are presented as follows:

$$\begin{aligned} \Delta y &= y_{pred} - y_{true} \\ L_1(\Delta y) &= \Delta y \\ L_2(\Delta y) &= (\Delta y)^2 \\ SmoothL_1(\Delta y) &= \begin{cases} 0.5\,(\Delta y)^2 & if\ |\Delta y| < 1 \\ |\Delta y| - 0.5. & else \end{cases} \end{aligned} \tag{6}$$

The total recognition loss and all location losses, including offset loss, size loss, and orientation loss, are presented by the following (5):

$$\begin{aligned} L_{total} &= L_{heat-MFL} + \frac{1}{N}\sum_{k=1}^{N} SmoothL_{1_{off}}(\hat{O}_k - O_k) \\ &+ \frac{1}{N}\sum_{k=1}^{N} SmoothL_{1_{size}}(\hat{S}_k - S_k) \\ &+ \frac{1}{N}\sum_{k=1}^{N} SmoothL_{1_{ori}}(\hat{\theta}_k - \theta_k) \end{aligned} \tag{7}$$

where $\hat{O}_k$, $\hat{S}_k$, and $\hat{\theta}_k$ are the predicted offset, size, and orientation, while O_k, S_k, and θ_k are the ground truth values of the offset, size, and orientation for object k, respectively. The proposed methods are evaluated and compared to previous 3D pedestrian detection approaches in the experimental section.

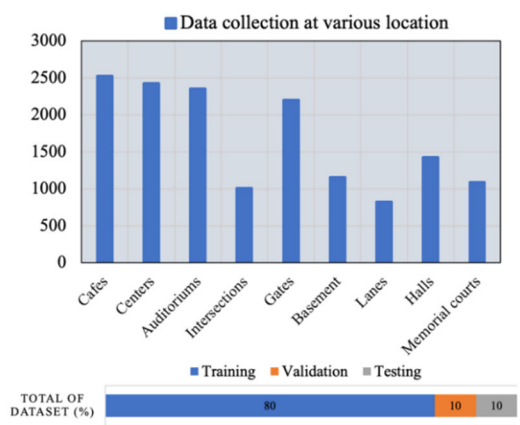

FIGURE 6. Data collection from various locations and data distribution.

IV. EXPERIMENTS

Our methodology utilized the PIXOR algorithm, a CNN-based 3D object detector tailored for processing LiDAR point clouds through a BEV representation. Central to

our experimental setup is the JRDB pedestrian detection dataset. Initially, we applied data augmentation techniques, such as rotation, scaling, and translation, to the JRDB dataset. This augmentation is vital to enhance the dataset's diversity and variability, fostering the development of a more robust and adaptable model. Subsequently, the augmented data undergoes a transformation into a BEV representation, followed by voxelization. The BEV representation offers a top-down view of the scene, crucial for effective pedestrian detection in LiDAR scans, while voxelization transforms this data into a 3D grid, streamlining processing. Finally, the processed data is used to train the IRBGHR-PIXOR model. This model is an enhanced version of the PIXOR algorithm, designed specifically for efficient and accurate pedestrian detection using 3D LiDAR point clouds.

A. DATASETS

The JRDB pedestrian detection dataset [14] was utilized for evaluating our proposed model. It comprises 15,000 samples, partitioned into training (80%), validation (10%), and testing (10%) sets. Data collection took place at various locations, including cafes, shopping centers, auditoriums, intersections,gates, basements, lanes, halls, and memorial courts, as illustrated in Fig. 6. The data distribution across these locations was different, categorizing crowd density into three levels: easy, moderate, and challenging, determined by the population density in each area.

B. DATA AUGMENTATION

Data augmentation is employed as a critical component for expanding the versatility and adaptability of our point cloud data. To ensure diversity, three fundamental augmentation techniques, namely "Rotation," "Scaling," and "Translation," are utilized. In the "Rotation" process, the data is subjected to a specified angle of rotation, introducing variety within an angular limit of $\mathbf{20}$ degrees. This approach enables the data to encompass a broad range of angular variations. For "Scaling," the goal is to encompass variations in object sizes and distances. The data is consistently scaled within a range of $\mathbf{0.95}$ ***to*** $\mathbf{1.05}$ times its original size, accommodating the diverse array of object scales. The "Translation" operation focuses on shifting the data to simulate different object locations. This is achieved by applying a random shift to the data and adjusting spatial positions using a scaling factor of $\mathbf{0.4}$. It's noteworthy that the augmentation process selects data randomly and effectively doubles the total number of samples, resulting in an overall dataset size of 30,000 samples. Fig.7 shows the normalization and augmentation of data.

C. EXPERIMENTS SETTING

The region of interest for the point cloud is set to $[-\mathbf{6.0}, \mathbf{6.0}] \times [-\mathbf{6}, \mathbf{6}]$ meters for the two x-y axes, and bird's eye view projection is carried out with a discretization resolution of $\mathbf{0.05}\boldsymbol{m}$. The height span is established as $[-\mathbf{1.5}, \mathbf{1.5}]$ meters within LiDAR coordinates to correspond with the typical human height, and all data points are segregated into 30 segments, each with a bin width of $\mathbf{0.1}\boldsymbol{m}$. In addition, a single reflectance channel is computed, resulting in our input representation having dimensions of $\mathbf{240} \times \mathbf{240} \times \mathbf{30}$. In contrast to other detectors [50] that commence by initializing network weights from a pre-trained model, we opt to train our network from the ground up, avoiding any reliance on pre-trained models.

The IRBGHR-PIXOR model is implemented in PyTorch [51]. A batch size of 10 is employed for training, which consists of 20 epochs. The model training is conducted on a system with Ubuntu 20.04, an Intel i7 3.4 GHz CPU, an Nvidia GTX 3060 GPU, and 32 GB of RAM. During training, the networks are optimized using a learning rate of $\mathbf{3}\times\mathbf{10}^{-\mathbf{4}}$. The optimization is carried out using the AdamW [52] optimizer utilizing a one-cycle policy [53]. A momentum of $\mathbf{0.9}$ is applied, and a fixed weight decay of $\mathbf{0.0005}$ is utilized to ensure convergence

D. EVALUATION METRIC

To assess the performance of our pedestrian detection model, we utilize two key evaluation metrics: Intersection over Union (IoU) and Average Precision (AP). IoU is defined as the ratio between the intersection and union of the predicted region and the actual object region. Here, "AoO" corresponds to the area of overlap, which is the intersecting area between the predicted and ground truth bounding boxes, while "AoU" encompasses the area of union, which combines the predicted bounding box and the ground truth. IoU values fall within the range (0,1), with each detection having a unique IoU score. To determine whether detection is correct or wrong, we employ a predefined threshold. This definition introduces the concepts of True Positive (TP), when IoU is above the threshold, representing correct detections; False Positive (FP), when IoU is below the threshold, signifying wrong detections; and False Negative (FN), which occurs when a ground truth object lacks a corresponding predicted bounding box. Additionally, in order to compute AP, precision and recall are defined, where precision measures the accuracy of predictions, and recall assesses the model's ability to identify correct detections. These metrics provide a comprehensive assessment of the model's performance in pedestrian detection. These definitions are expressed in the following equations:

$$IoU = \frac{Area\ of\ Overlap\ (AoO)}{Area\ of\ Union\ (AoU)} \tag{8}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}} = \frac{\boldsymbol{TP}}{\boldsymbol{All\ of\ Predicted\ box}} \tag{9}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}} = \frac{\boldsymbol{TP}}{\boldsymbol{All\ of\ ground\ truth\ box}} \tag{10}$$

$$\boldsymbol{AP} = \int_{\mathbf{0}}^{\mathbf{1}} \boldsymbol{Precision}\ (\boldsymbol{Recall})\boldsymbol{d}(\boldsymbol{Recall}) \tag{11}$$

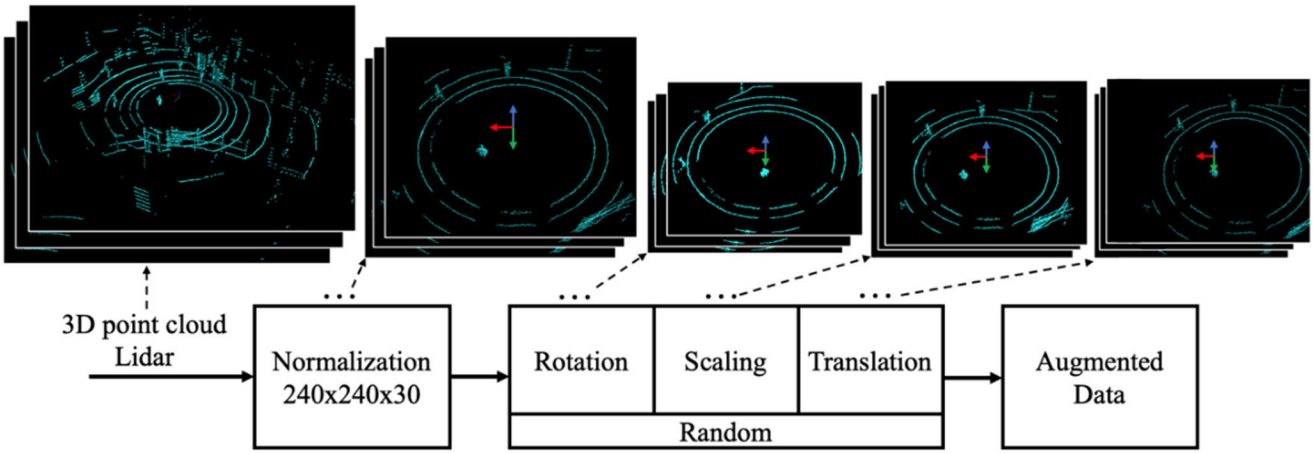

FIGURE 7. Normalization and augmentation of data throughout the training process.

TABLE 1. Comparative evaluation and accuracy assessment of 3D pedestrian detection systems on the JRDB test set with varying IoU thresholds and interest ranges.

Backbone Network	Avg_Training Time per Epoch (Seconds)	$AP_{IoU=0.5}$, test (%)			$AP_{IoU=0.7}$, test (%)			$AP_{IoU=0.9}$, test (%)		
		0-6m	6-12m	12-24m	0-6m	6-12m	12-24m	0-6m	6-12m	12-24m
PIXOR without GAU [8]	1262	87.38	73.16	63.53	85.51	70.35	58.42	64.63	51.58	40.65
PIXOR + GAU [8]	1320	90.52	73.63	64.62	88.73	71.94	62.41	67.92	52.15	42.55
AFDet + GAU [47]	3649	95.64	76.39	65.05	94.84	75.28	62.58	76.11	**54.81**	**44.23**
IRBGHR-PIXOR + GAU	**993**	**97.17**	**79.74**	**66.55**	**95.27**	**77.15**	**63.17**	**77.48**	54.47	43.86

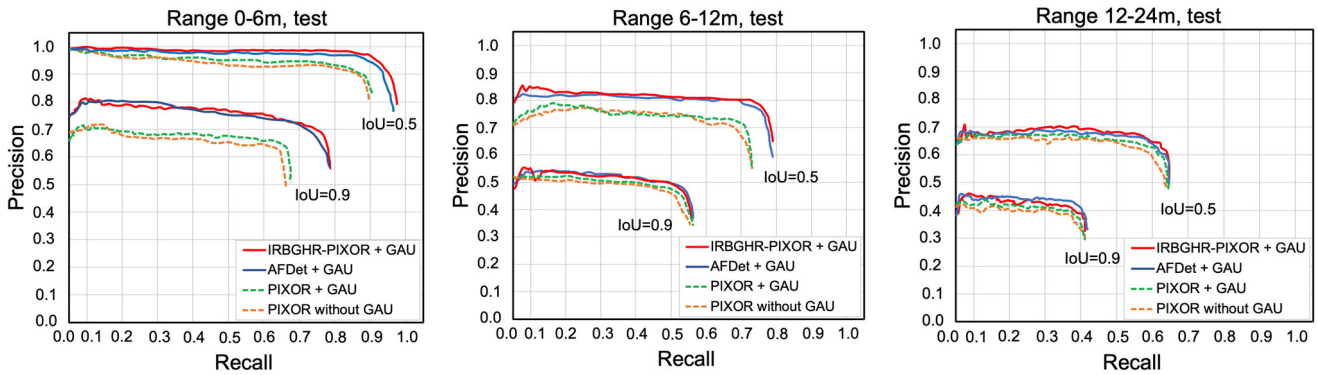

FIGURE 8. Evaluation of proposed and conventional methods on JRDB test set: analysis of 3 regions of interest at IoU=0.5 and high IoU = 0.9 levels.

E. EVALUATION RESULTS

1) IMPACT OF LOSS FUNCTIONS

In our experiments, we conducted different types of comparisons involving our proposed model and traditional methods like PIXOR [8] and AFDet [54] combined with a Gaussian heatmap. The evaluation results are presented in TABLE 1. From the table, it is evident that the proposed approach outperforms other conventional methods in terms of accuracy, with AP scores exceeding those of other approaches at IoU levels of $\mathbf{0.5, 0.7,}$ and $\mathbf{0.9}$, within a range of 6 meters, which aligns with the requirements of indoor robots. Notably, the proposed method exhibits an approximately 10% improvement over the conventional PIXOR approach. However, when compared to AFDet at IoU $= \mathbf{0.9}$ within the distance ranges of $\mathbf{6}\boldsymbol{m} - \mathbf{12}\boldsymbol{m}$ and $\mathbf{12}\boldsymbol{m} - \mathbf{24}\boldsymbol{m}$, it lags behind by less than $\mathbf{1\%}$. Nevertheless, it's important to note that the training speed of the proposed method is four times slower for a single epoch. Moreover, given that the robot operates exclusively indoors, the requirement for high accuracy at shorter distances (below $\mathbf{6}\boldsymbol{m}$) and a negligible performance difference at longer distances (with $< \mathbf{1\%}$ error) is acceptable, considering its advantages.

TABLE 2. Impact of loss function on accuracy.

Recognition	Localization	$AP_{\text{IoU=0.5}}$, *val* (0-6m)	$AP_{\text{IoU=0.5:0.9}}$, *val* (0-6m)
Cross-entropy	L1	**91.63%**	**82.56%**
Cross-entropy	Smooth L1	**92.45%**	**83.05%**
Focal	L1	**93.89%**	**85.03**
Focal	Smooth L1	94.25%	**85.16%**
Modified-FL	L1	96.25%	**86.72%**
Modified-FL	**Smooth L1**	**97.21%**	**88.30%**

The fine-grained Precision-Recall (PR) curves are shown for the four methods at IOU levels of 0.5 and 0.9 across three different distance thresholds in Fig. 8. The curve plots once again demonstrate the superior accuracy of our proposed approach compared to the pedestrian detection capabilities of the other three methods at close range. The significant improvement in system accuracy in detecting pedestrians for 3D point cloud data is underscored by the substantial impact of both the proposed backbone replacement and the application of Gaussian heatmap regression (GAU) for heat recognition. Remarkably, the integration of the Gaussian heatmap has demonstrated a pronounced efficacy, notably in its capacity to augment the recognition of diminutive objects set against an expansive spatial backdrop. This augmentation is conspicuously reflected in the accuracy metrics when contrasting the conventional approach devoid of Gaussian heatmap with its Gaussian heatmap-equipped counterpart, yielding an approximate 3% increase in accuracy.

Subsequently, an assessment of the loss functions introduced for both recognition and localization heads is performed. The results, depicted in TABLE 2, underscore the model's training procedure, encompassing the integration of a range of loss functions, such as Cross-entropy, Focal loss, and Modified Focal loss, in combination with L1 loss and Smooth L1 loss. Following this, a performance evaluation on the validation dataset is executed, comparing outcomes based on two critical accuracy metrics ($AP_{\text{IoU}} = 0.5$ and $AP_{\text{IoU=0.5:0.9}}$). The use of Cross-entropy and L1 loss, while respectable, yields an $AP_{\text{IoU=0.5}}$ of 91.63% and an $AP_{IoU=0.5:0.9}$ of 82.56%. Incorporating Cross-entropy with Smooth L1 loss brings an improvement, resulting in $AP_{IoU=0.5}$ of 92.45% and $AP_{IoU=0.5:0.9}$ of 83.05%. Meanwhile, the employment of Focal loss alongside L1 loss introduces a notable performance boost, achieving an $AP_{IoU=0.5} of\, 93.89\%$ and an $AP_{IoU=0.5:0.9}$ of 85.03%. Furthermore, combining Focal loss with Smooth L1 loss enhances accuracy, resulting in an $AP_{IoU=0.5}$ of 94.25% and an $AP_{IoU=0.5}$ of 85.16%.

Notably, the pinnacle of performance is attained with the use of Modified Focal loss and L1 loss, achieving an $AP_{IoU=0.5}$ of 96.25% and an $AP_{IoU=0.5:0.9}$ of 86.72%. The culmination of these advancements is witnessed in the proposed method, where Modified Focal loss harmonizes with Smooth L1 loss to achieve peak accuracy, boasting an $AP_{IoU=0.5}$ of 97.21% and an $AP_{IoU=0.5:0.9}$ of 88.30%. This comprehensive evaluation underscores the remarkable strides taken in pedestrian detection, with the "Modified-FL" and "Smooth L1" combination emerging as the superior choice. The "Modified-FL" loss function emerges as a robust solution, effectively handling challenging scenarios while also addressing data imbalance, distances, and probability values of ground truth, thereby enhancing the overall performance of the detection system. In parallel, "Smooth L1" proves its mettle in comparison to the traditional "L1" loss by offering greater resilience to outliers, contributing to training stability, and ensuring precise localization. The selection of these adept loss functions is pivotal in advancing the accuracy and reliability of pedestrian detection in 3D point cloud data.

TABLE 3. Impact of alpha (α) and beta (β) parameters on pedestrian detection accuracy for modified-focal loss.

IRBGHR-PIXOR + GAU (Modified-Focal)	$AP_{\text{IoU=0.5}}$, *val* (0-6m)	$AP_{\text{IoU=0.5:0.9}}$, *val* (0-6m)
$\alpha = 1, \beta = 1$	94.77%	85.03%
$\alpha = 1, \beta = 2$	95.27%	86.22%
$\alpha = 1, \beta = 3$	96.23%	87.12%
$\alpha = 1, \beta = 4$	96.72%	87.43%
$\alpha = 1, \beta = 5$	96.47%	87.28%
$\alpha = 2, \beta = 1$	95.36%	86.66%
$\alpha = 2, \beta = 2$	96.41%	87.07%
$\alpha = 2, \beta = 3$	96.89%	87.57%
$\alpha = 2, \beta = 4$	97.21%	88.30%
$\alpha = 2, \beta = 5$	97.05%	87.42%

2) PARAMETER TUNING

Following the evaluation of the proposed loss function, which demonstrated a significant enhancement in accuracy, we proceeded to experimentally determine the optimal values for the alpha and beta weights within the recognition loss function. The results of this evaluation, involving various combinations of α and $\boldsymbol{\beta}$ values in (5), are presented in TABLE 3. The observed results exhibit a consistent upward trend in performance with the incremental increase of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ values. The observed results exhibit a consistent upward trend in performance with the incremental increase of both alpha and beta values. Notably, the highest performance is achieved with higher alpha and beta values, such as $\boldsymbol{\alpha} = 2$ and $\boldsymbol{\beta} = 4$, which yield the most favorable $AP_{IoU=0.5}$ and $AP_{IoU=0.5:0.9}$ scores at 97.21% and 88.30%, respectively. These outcomes underscore the critical importance of meticulous parameter tuning, particularly concerning alpha and beta, to optimize the recognition loss function. This optimization process profoundly influences the overall accuracy of the pedestrian detection system. It's crucial to highlight that augmenting the beta values imposes a more substantial penalty within the loss function, particularly when dealing with ground truth values smaller than 1. This penalization mechanism plays a

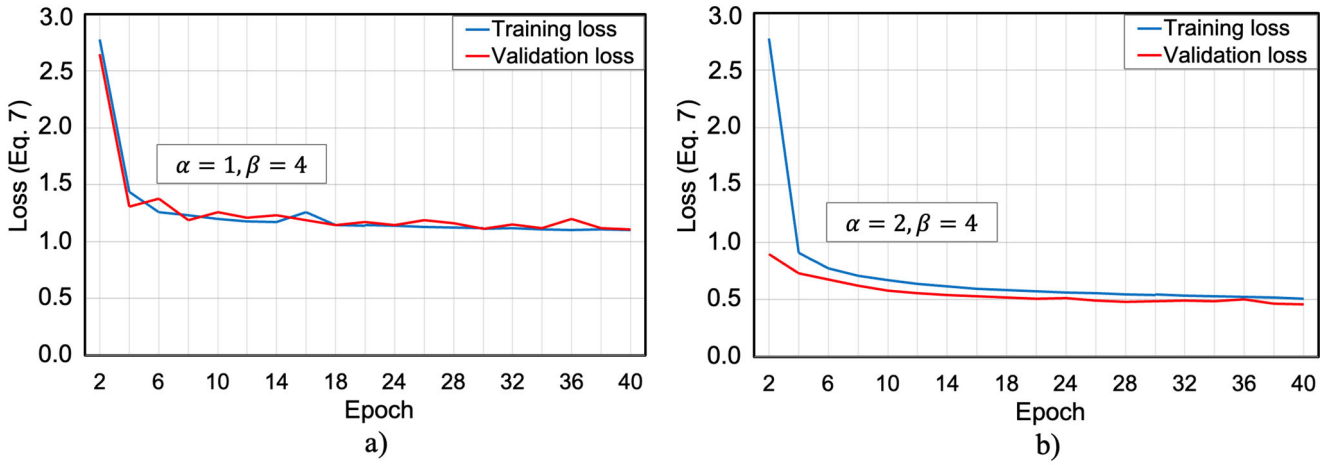

FIGURE 9. Effect of α and β parameters on loss function convergence. a) with $\alpha = 1$, $\beta = 4$, b) with $\alpha = 2$ and $\beta = 4$.

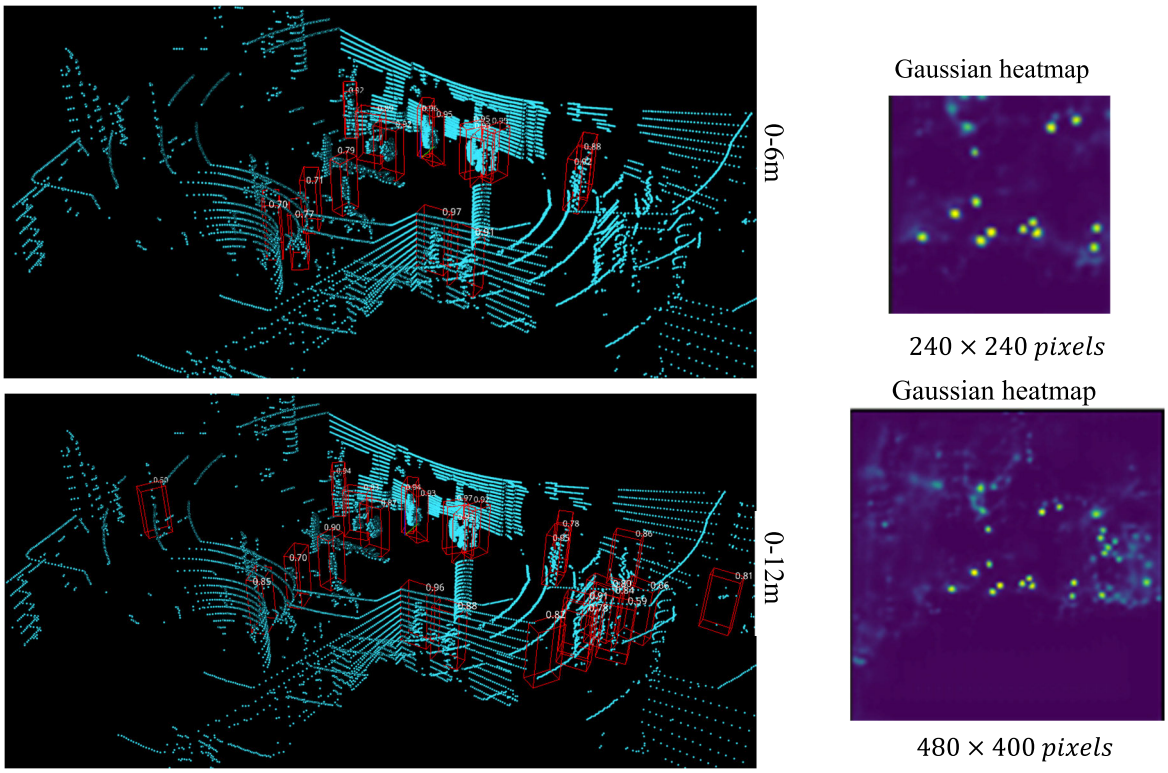

FIGURE 10. Pedestrian detection results at an intersection within 0-6m and 0-12m ranges of interest with sparse pedestrian level.

pivotal role in improving the model's capacity to recognize pedestrians, especially in situations involving smaller objects or challenging detection scenarios.

Fig. 9 vividly illustrates the trajectory of the loss function values, encapsulating both Training Loss and Validation Loss, for two distinct alpha and beta parameter combinations: $\boldsymbol{\alpha} = \mathbf{1}$, $\boldsymbol{\beta} = \mathbf{4}$, and $\boldsymbol{\alpha} = \mathbf{2}$, $\boldsymbol{\beta} = \mathbf{4}$ using the training set of the dataset. An intriguing trend comes to the forefront, highlighting the notably swift convergence achieved when employing $\boldsymbol{\alpha} = \mathbf{2}$ and $\boldsymbol{\beta} = \mathbf{4}$ throughout a 40-epoch training phase. It's noteworthy that the $\boldsymbol{\alpha} = \mathbf{2}$ and $\boldsymbol{\beta} = \mathbf{4}$ combination reaches a remarkable minimum loss function value of approximately 0.5, contrasting with the higher value of around $\mathbf{1.2}$ for the counterpart coefficient pair. This insight underscores the heightened efficiency that $\boldsymbol{\alpha} = \mathbf{2}$ and $\boldsymbol{\beta} = \mathbf{4}$ impart to the training process. The model's ability to swiftly converge within a training period of 40 epochs emphasizes the compelling advantages of fine-tuning these parameters for an expedited and highly effective training regimen in the realm of pedestrian detection using 3D point cloud data.

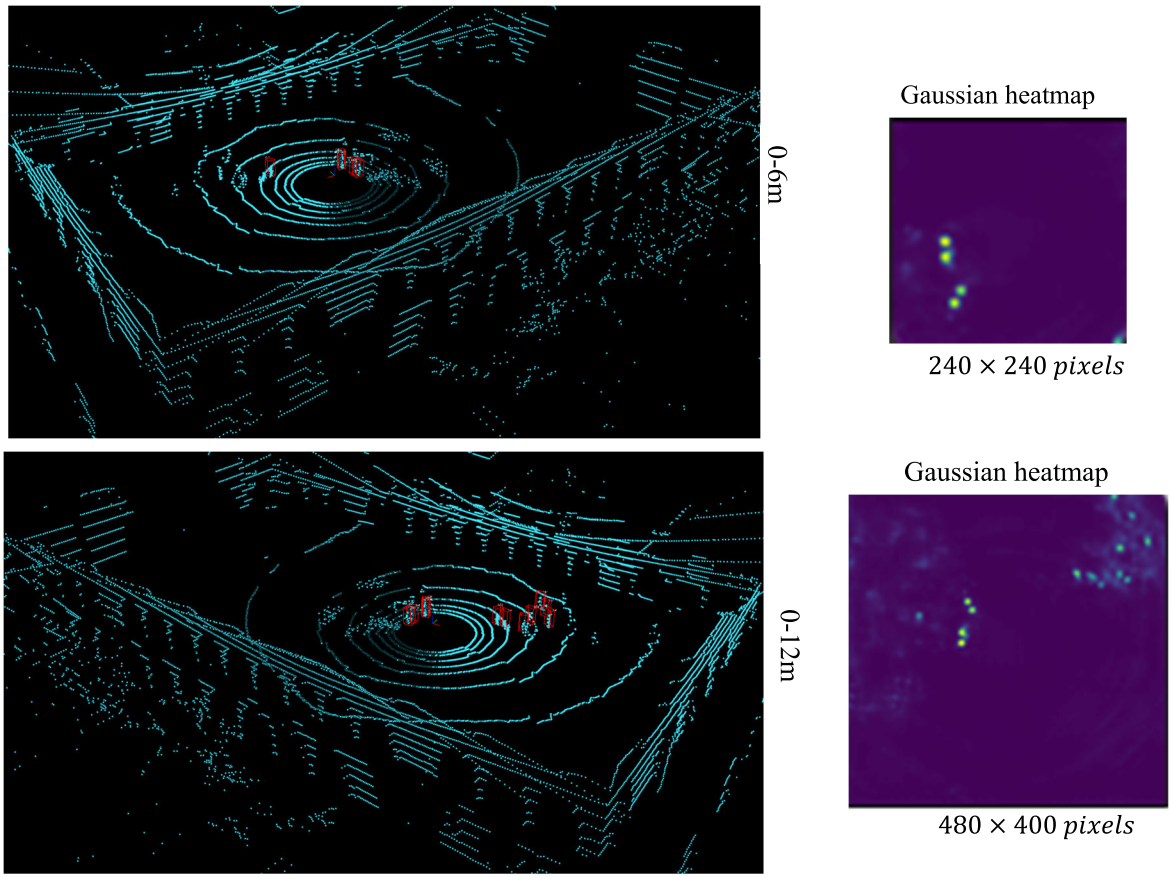

FIGURE 11. Pedestrian detection results in a center within 0-6m and 0-12m ranges of interest with sparse pedestrian level.

TABLE 4. Impact of data augmentation on proposed method.

	Data Aug.	$AP_{\text{IoU}=0.5}$, *val* (0-6m)	$AP_{IoU=0.5:0.9}$, *val* (0-6m)
IRBGHR-PIXOR + GAU (Modified-Focal)	none	**91.06%**	**83.42%**
	Rotation + Scaling + Translation	**97.21%**	**88.30%**

3) IMPACT OF DATA AUGMENTATION

TABLE 4 presents a comparative evaluation of the impact of data augmentation on the accuracy of the IRBGHR-PIXOR + GAU (Modified-Focal) method for pedestrian detection. Two augmentation scenarios are considered: one with no augmentation and another involving Rotation, Scaling, and Translation. The results clearly demonstrate the substantial effect of data augmentation on model performance. When no augmentation is applied, the model achieves an $AP_{IoU=0.5}$ of **91.06%** and an $AP_{IoU=0.5:0.9}$ of **83.42%.** However, the introduction of Rotation, Scaling, and Translation significantly enhances accuracy, with a notable boost in $AP_{IoU=0.5}$ to **97.21%** and $AP_{IoU=0.5:0.9}$ to 88.30%. This observation underscores the pivotal role of data augmentation in enhancing the model's capability to detect pedestrians effectively. By introducing variations in the training data through these transformations, the model becomes more robust, improving its accuracy across a broader range of real-world scenarios. It signifies the practical importance of incorporating data augmentation techniques to bolster the reliability and precision of pedestrian detection in the context of 3D point cloud data.

In the final segment, the outcomes of pedestrian detection in real-world data will be showcased, encompassing diverse locations, spanning from sparsely populated, uncomplicated areas with sparse pedestrian presence to intricate, densely populated environments replete with obstacles, as depicted in Fig. 10. 11., and 12. Additionally, the predicted Gaussian heatmap is also depicted in the figure.

F. DISCUSSION

The backbone architecture refinements directly enable a 6.5% boost in Average Precision (AP) for pedestrian detection (from 73.16% to 79.74% at 0.5 IOU). This highlights the value of multi-scale feature learning using inverted residual blocks in processing complex spatial LiDAR data.

Additionally, integrating Gaussian heatmap regression improves AP by 3-4% over baseline across IOU thresholds.

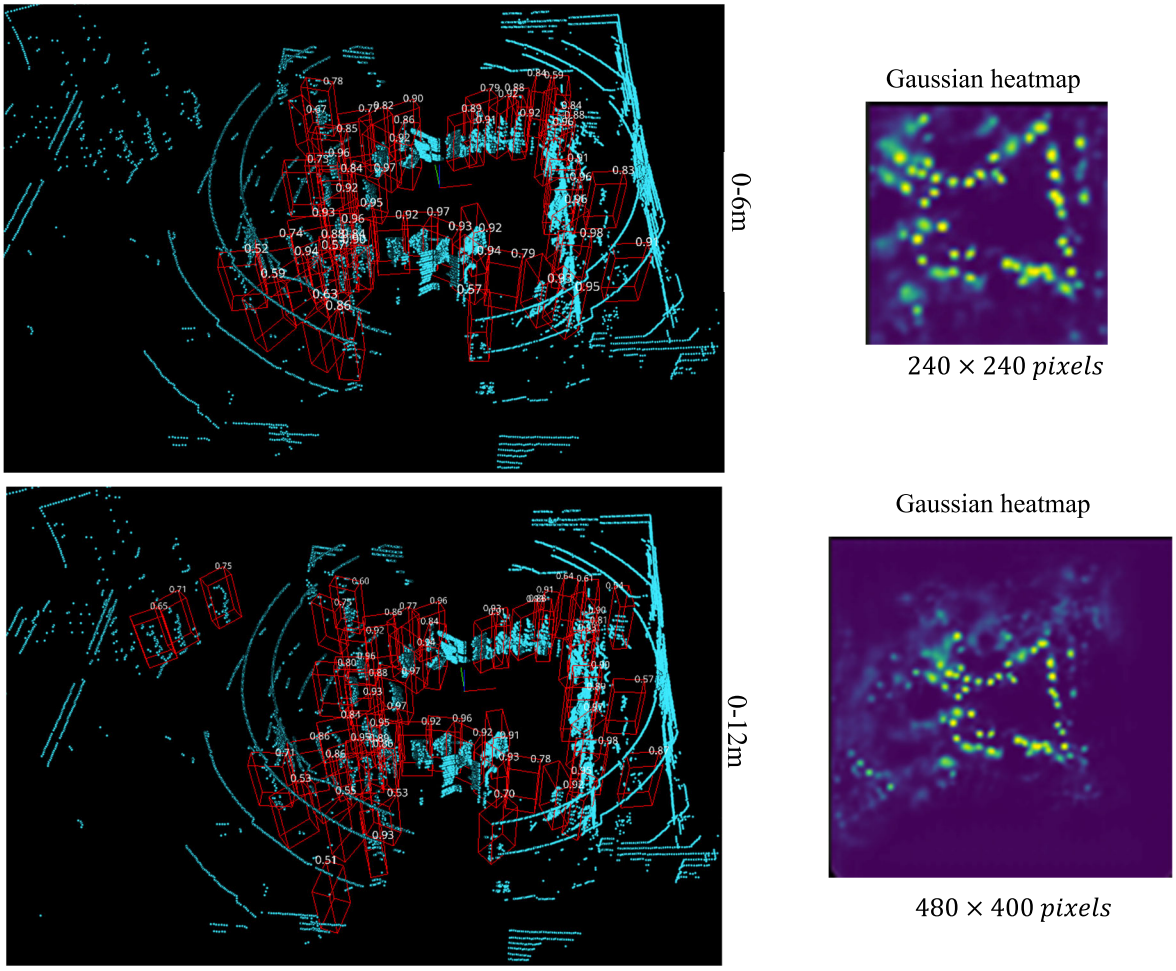

FIGURE 12. Pedestrian detection results at a hall within 0-6m and 0-12m ranges of interest with high pedestrian density.

This showcases its ability to handle variability in human shapes/sizes and overcome sensor noise/occlusions.

Furthermore, optimizing the modified focal loss function parameters results in a 7% increase in AP over standard focal loss. Finding the optimal α, β hyperparameter settings is vital for emphasizing precise localization while mitigating class imbalance.

Cumulatively, the architectural and algorithmic improvements in IRBGHR-PIXOR advance state-of-the-art to 97.17% AP on JRDB. The 5-10% gains over existing LiDAR detectors validate the solutions for enhancing indoor robotic perception.

The targeted enhancements directly address key challenges that emerge when adapting pedestrian detection for unstructured indoor settings. This research provides an effective template to boost reliability using LiDAR for robotic applications.

V. CONCLUSION

This paper presented a novel 3D pedestrian detection system called IRBGHR-PIXOR tailored for indoor robotics applications. The key contributions include:

First, an improved PIXOR backbone using inverted residual blocks and feature fusion to enhance multi-scale feature learning for small objects.

Second, integration of Gaussian heatmap regression in the detection head to precisely localize pedestrian keypoints against complex backgrounds.

Third, a modified focal loss function to handle class imbalance while emphasizing precise localization.

Fourth, extensive experiments on the indoor JRDB dataset demonstrated state-of-the-art accuracy of 97.17% AP at 0.5 IOU, outperforming prior methods.

The proposed architectural improvements and training strategies significantly advance the state-of-the-art in LiDAR-based pedestrian detection for robots operating in indoor spaces. Precise 3D perception of humans allows safe navigation and robust scene understanding in dynamic, unstructured environments.

Several promising avenues exist to build on this research and further advance reliable pedestrian detection for robotics. More advanced augmentation techniques like generative adversarial networks (GANs) could produce highly realistic synthetic data to improve model robustness across indoor

scenes. Additionally, adaptive loss tuning approaches based on meta-learning must be explored for scene-adaptive parameter optimization instead of manual tuning. Rigorously evaluating performance on diverse benchmark datasets would provide invaluable insight into generalizability across varying crowd densities, illumination conditions and sensor noise profiles. Fusing LiDAR with complementary RGB and depth data from vision sensors can enrich representations through multi-modal fusion to exploit synergies between active and passive sensing. Testing long-term real-world functionality would reveal reliability challenges in complex deployments, guiding the development of online domain adaptation techniques. Thoroughly investigating these areas of more advanced augmentation, adaptive loss tuning, multi-dataset evaluation, sensor fusion and online adaptation would significantly elevate pedestrian detection to overcome limitations in robustness, adaptability and contextual reasoning. This research has established the foundations to progress these multifaceted efforts toward deployment-ready perception.

ACKNOWLEDGMENT

The authors would like to thank the support of time and facilities from the Ho Chi Minh City University of Technology (HCMUT), e Vietnam National University Ho Chi Minh City (VNU-HCM), and FPT University, Can Tho, Vietnam, for this study.

REFERENCES

[1] R. Martín-Martín, M. Patel, H. Rezatofighi, A. Shenoi, J. Gwak, E. Frankel, A. Sadeghian, and S. Savarese, "JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6748–6765, Jun. 2023.

[2] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3D object detection methods for autonomous driving applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3782–3795, Oct. 2019.

[3] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection," *Int. J. Comput. Vis.*, vol. 131, no. 2, pp. 531–551, Feb. 2023.

[4] L. Wang and Y. Huang, "A survey of 3D point cloud and deep learning-based approaches for scene understanding in autonomous driving," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 6, pp. 135–154, Nov. 2022.

[5] Z. Yin and T. Oncel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[6] H. Wang, Z. Chen, Y. Cai, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "Voxel-RCNN-complex: An effective 3-D point cloud object detector for complex traffic conditions," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[7] H. L. Alex, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12697–12705.

[8] B. Yang, W. Luo, and R. Urtasun, "PIXOR: Real-time 3D object detection from point clouds," 2019, *arXiv:1902.06326*.

[9] Y. Wu, Y. Wang, S. Zhang, and H. Ogai, "Deep 3D object detection networks using LiDAR data: A review," *IEEE Sensors J.*, vol. 21, no. 2, pp. 1152–1171, Jan. 2021.

[10] Y. Li, L. Ma, Z. Zhong, F. Liu, M. A. Chapman, D. Cao, and J. Li, "Deep learning for LiDAR point clouds in autonomous driving: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3412–3432, Aug. 2021.

[11] M. S. Mekala, W. Park, G. Dhiman, G. Srivastava, J. H. Park, and H.-Y. Jung, "Deep learning inspired object consolidation approaches using LiDAR data for autonomous driving: A review," *Arch. Comput. Methods Eng.*, vol. 29, no. 5, pp. 2579–2599, Aug. 2022.

[12] R. Q. Charles, L. Wei, W. Chenxia, S. Hao, and J. G. Leonidas, "Frustum PointNets for 3D object detection from RGB-D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 918–927.

[13] M. Ehsanpour, F. Saleh, S. Savarese, I. Reid, and H. Rezatofighi, "JRDB-act: A large-scale dataset for spatio-temporal action, social group and activity detection," 2021, *arXiv:2106.08827*.

[14] E. Vendrow, D. Tho Le, J. Cai, and H. Rezatofighi, "JRDB-pose: A large-scale dataset for multi-person pose estimation and tracking," 2022, *arXiv:2210.11940*.

[15] G. Andreas, L. Philip, and U. Raquel, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[16] L. Zhengxiong, W. Zhicheng, H. Yan, T. Tan, and Z. Erjin, "Rethinking the heatmap regression for bottom-up human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13264–13273.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[19] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. L. Cun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–16.

[20] J. Pont-Tuset, P. Arbeláez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, Jan. 2017.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[22] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[25] S. Huang, L. Liu, X. Fu, J. Dong, F. Huang, and P. Lang, "Overview of LiDAR point cloud target detection methods based on deep learning," *Sensor Rev.*, vol. 42, no. 5, pp. 485–502, Aug. 2022.

[26] D. Fernandes, A. Silva, R. Névoa, C. Simoes, D. Gonzalez, M. Guevara, P. Novais, J. Monteiro, and P. Melo-Pinto, "Point-cloud based 3D object detection and classification methods for self-driving applications: A survey and taxonomy," *Inf. Fusion*, vol. 68, pp. 161–191, Apr. 2021.

[27] M. Arsalan, A. Dragomir, F. John, F. John, and K. Jana, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 7074–7082.

[28] R. Q. Charles, S. Hao, K. Mo, and J. G. Leonidas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 652–660.

[29] R. Q. Charles, Y. Li, S. Hao, and J. G. Leonidas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–10.

[30] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, p. 3337, Oct. 2018, doi: 10.3390/s18103337.

[31] W. Shi and R. Rajkumar, "Point-GNN: Graph neural network for 3D object detection in a point cloud," 2020, *arXiv:2003.01251*.

[32] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li, "PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection," 2021, *arXiv:2102.00463*.

[33] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," 2019, *arXiv:1912.13192*.
[34] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," 2019, *arXiv: 1911.10150*.
[35] O. Rinchi, H. Ghazzai, A. Alsharoa, and Y. Massoud, "LiDAR technology for human activity recognition: Outlooks and challenges," *IEEE Internet Things Mag.*, vol. 6, no. 2, pp. 143–150, Jun. 2023.
[36] C. Xiaozhi, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals for accurate object class detection," in *Proc. NIPS*, 2015, pp. 1–9.
[37] V. A. Christine, S.-N. Jaya, and S.-N. Jaya, "Building 3D virtual worlds from monocular images of urban road traffic scenes," in *Advances in Visual Computing* (Lecture Notes in Computer Science). 2021.
[38] M. Zhu, S. Zhang, Y. Zhong, P. Lu, H. Peng, and J. Lenneman, "Monocular 3D vehicle detection using uncalibrated traffic cameras through homography," 2021, *arXiv:2103.15293*.
[39] Y. Chen, F. Liu, and K. Pei, "Monocular vehicle 3D bounding box estimation using homograhy and geometry in traffic scene," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1995–1999.
[40] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau, "Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image," 2017, *arXiv:1703. 07570*.
[41] L. Lijie, J. Lu, C. Xu, Q. Tian, and J. Zhou, "Deep fitting degree scoring network for monocular 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1057–1066.
[42] Y. Chen, S. Liu, X. Shen, and J. Jia, "DSGN: Deep stereo geometry network for 3D object detection," 2020, *arXiv: 2001.03398*.
[43] A. Simonelli, S. R. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3D object detection," 2019, *arXiv:1905.12365*.
[44] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.
[45] V. S. Hikkal, "Comparative study of 3D object detection frameworks based on LiDAR data and sensor fusion techniques," *J. Phys., Conf.*, vol. 2232, May 2022, Art. no. 012015.
[46] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[47] P. Ding, H. Qian, and S. Chu, "SlimYOLOv4: Lightweight object detector based on YOLOv4," *J. Real-Time Image Process.*, vol. 19, no. 3, pp. 487–498, Jun. 2022.
[48] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 642–656, Mar. 2020.
[49] M. Weber, M. Fürst, and J. M. Zöllner, "Automated focal loss for image based object detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 1423–1429.
[50] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 1907–1915.
[51] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," Tech. Rep., 2017.
[52] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–19.
[53] S. Gugger. (2018). *The 1Cycle Policy*. [Online]. Available: https://sgugger.github.io/the-1cycle-policy.html
[54] R. Ge, Z. Ding, Y. Hu, Y. Wang, S. Chen, L. Huang, and Y. Li, "Afdet: Anchor free one stage 3D object detection," 2020, *arXiv: 2006.12671*.

DUY ANH NGUYEN received the bachelor's degree in mechatronic engineering from the Ho Chi Minh City University of Technology (HCMUT), Vietnam, in 2015, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2020. He was a Postdoctoral Researcher with the Korea Advanced Institute of Science and Technology (KAIST), until 2023. He is currently a Lecturer with HCMUT. He is also a Researcher and an Educator in the field of mechatronic engineering and related disciplines. His research interests include frequency comb, photonics, computer vision, and intelligent systems.

KHANG NGUYEN HOANG is currently pursuing the degree in software engineering with the Department of Software Engineering, FPT University, Can Tho, Vietnam.

His research interests include machine learning, deep learning, autonomous robotics, and image processing.

NGUYEN TRUNG NGUYEN is currently pursuing the degree in software engineering with the Department of Software Engineering, FPT University, Can Tho, Vietnam.

His research interests include autonomous robots, computer vision, deep learning, and machine learning.

DUY ANH NGUYEN received the bachelor's degree in automatic control from the Ho Chi Minh City University of Technology (HCMUT), Vietnam, in 2003, and the master's and Ph.D. degrees in logistics from Korea Maritime University, in 2006 and 2009, respectively. He is currently an Associate Professor with the Faculty of Mechanical Engineering, HCMUT. He is also a Researcher and an Academician in the fields of mechatronic engineering, automation, robotics, and logistics. His research interests include logistics, automation, robotics, mechatronics, computer vision, and manufacturing technologies.

HOANG NGOC TRAN received the B.S. degree in mechatronics engineering from the Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam, in 2015, and the Ph.D. degree in electrical and computer engineering from Sungkyunkwan University, Suwon, South Korea, in 2020.

From 2020 to 2022, he was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Sungkyunkwan University. Since 2022, he has been a Lecturer and a Researcher with the Department of Software Engineering, FPT University, Can Tho, Vietnam. His research interests include signal processing, motion control, embedded systems, autonomous robotics, computer vision, machine learning, and deep learning.