

Li2Former: Omni-Dimension Aggregation Transformer for Person Detection in 2-D Range Data

Haodong Yang¹, Yadong Yang¹, Chenpeng Yao, Chengju Liu¹, and Qijun Chen¹, *Senior Member, IEEE*

Abstract—The sparsity of plane point clouds remains human detection based on 2-D LiDAR a challenging task. Recent deep-learning-based methods have made significant progress, but their unidimensional feature modeling hinders the flow of information across different axes (i.e., temporal, spatial, and batched). Besides, convolution operators limit the effective receptive field (ERF) of the model. To promote multidimensional interaction, we propose an omni-dimensional aggregation (ODA) module composed of a main aggregation stream and an auxiliary feature stream to assist in the temporal-spatial joint encoding as well as implicitly models the potential relationships between samples. By cascading multiple ODA blocks, the semantic gaps between different features are gradually bridged. In addition, to improve network convergence and generalization, an adaptive focus mechanism (AFM) is designed to guide the model to optimize more on difficult samples instead of simple or low-quality samples. Extensive experiments demonstrate the effectiveness of the proposed pipeline, achieving 76.4% and 81.4% AP on the DROW and JRDB benchmark datasets, respectively, surpassing the existing state-of-the-art (SOTA) results. Competitive results are also achieved in inference speed. Furthermore, we successfully validate the proposed model through real-world environments. Our code is available at <https://github.com/ai-winter/Li2Former-v1>.

Index Terms—2-D LiDAR, deep learning, human detection and tracking.

I. INTRODUCTION

IN HUMAN-ROBOT interaction scenarios (e.g., health care and security monitor), precise detection of the pedestrians helps robot understand the information related to human intentions, preferences, and more, thus improving the robot's service level [1], [2]. Due to the limited field of view and inaccurate distance measurement of RGB(-D) cameras [3], and the high-computational and storage costs of 3-D LiDAR [4], 2-D LiDAR has become a balanced

choice between performance and cost. However, the sparsity caused by the single horizontal scanning line of 2-D LiDAR presents difficulties in feature extraction, making reliable human detection still a challenging task.

Mining human-related information from sparse range measurements is important for achieving person detection. Traditional methods rely on precisely designed handcrafted features for template matching [5], [6], [7], [8], [9], but complex situations in the real world such as occlusions caused by human social interactions and numerical fluctuations resulting from ego motion usually render these predesigned approaches ineffective.

With the development of deep learning, implicitly extracting optimal human features from sparse point clouds can achieve higher accuracy and robustness than conventional methods. However, the existing works have the following weaknesses.

- 1) *Unidimensional Aggregation Operators*: Focusing only on spatial [10], [11] or temporal features [12], [13] results in insufficient information fusion, while ignoring explicit use of batch dimension. Recent studies [14], [15] indicate that the correlation between samples is equally crucial for improving data imbalance and enhancing the model generalization.
- 2) *Local Convolutional Operators*: Planar point clouds are essentially 1-D depth sequences containing global representations, such as object contours and shape descriptions [16]. Effective receptive field (ERF) limitation caused by local feature modeling is inefficient in capturing long-distance dependencies.

In this work, an approach named LiDAR-based 2-D person detection Transformer (Li2Former) is proposed. To overcome constraint 1), as shown in Fig. 1, the omni-dimensional aggregation (ODA) module is designed to extend the unidimensional aggregation operator to encompass the full axes across temporal, spatial, and batched dimensions. In contrast to the single-stream pipeline concatenated with various dimensional encoders, a novel two-stream training strategy is designed to progressively address the semantic gaps across omni-axis, where the main aggregation stream models the temporal features of point clouds and the auxiliary feature stream conducts spatial structural learning and internal sample data augmentation. Moreover, the auxiliary feature stream can be removed during the inference stage, which avoids the inconsistency between training and testing [17] as well as the additional computational burden. To address shortcoming 2), Transformer-based encoders are introduced to model the

Manuscript received 18 January 2024; revised 24 May 2024; accepted 12 June 2024. Date of publication 28 June 2024; date of current version 18 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62173248, Grant 62333017, and Grant 62233013; in part by Suzhou Key Industry Technological Innovation-Core Technology Research and Development Program under Grant SGC2021035; in part by Special Funds for Jiangsu Science and Technology Plan under Grant BE2022119; in part by Suzhou Key Research and Development Project under Grant SGC2021069; and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100. The Associate Editor coordinating the review process was Dr. Bruno Andò. (Corresponding authors: Qijun Chen; Chengju Liu.)

The authors are with the Department of Electronics and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: 2230700@tongji.edu.cn; 22310076@tongji.edu.cn; liuchengju@tongji.edu.cn; qjchen@tongji.edu.cn).

Digital Object Identifier 10.1109/TIM.2024.3420353

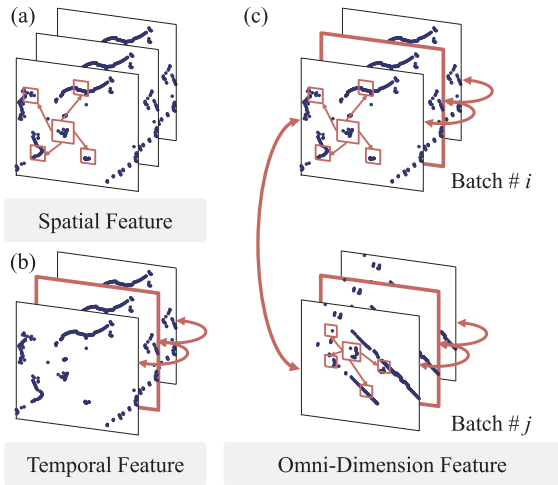


Fig. 1. ODA module promotes the fusion of features along the spatial, temporal, and batch axes, providing more comprehensive information interaction than unidimensional aggregation operators. (a) Spatial information from a single-frame point cloud. (b) Temporal information from consecutive frame point clouds. (c) Fusion of features along the spatial, temporal and batch axes, which provides more comprehensive information interaction than unidimensional aggregation operators.

ODA, enlarging the ERF through global attention calculation. Besides, considering the slow convergence of Transformer-based model, an adaptive focus mechanism (AFM), based on the geometric metric of predicted ground-truth centroid Gaussian distance, is proposed. In comparison with the similar convergence rate of traditional IoU regression across all samples, AFM applies greater gradient boosts to challenging samples, while reducing learning weights for simple or low-quality invalid samples, guiding the network to more effectively learn internal data patterns.

In summary, our main contributions are threefold.

- 1) Propose a Transformer-based ODA module, which injects spatial features of point clouds and sample correlations from the auxiliary feature stream into temporal information of continuous scans in the main aggregation stream, for omni-axes feature fusion and encoding.
- 2) Design a Gaussian-based AFM, which provides gain to the learning gradient of difficult samples, guiding the model in learning high-quality features. Experimental results demonstrate that the proposed method surpasses the previous state-of-the-art (SOTA) detection accuracy.
- 3) Release the code for PyTorch-based pedestrian detection models, and provide user-friendly ROS nodes to facilitate further research in robot applications.

The remainder of this article is structured as follows. Section II provides an overview about related work. Section III starts with the problem formulation for 2-D person detection, followed by an outline of the overall system. Then, the details about the proposed Li2Former pipeline are presented. Section IV gives simulation and physical experiments to validate the method, and we draw some conclusions in Section V.

II. RELATED WORK

A. Human Detection From 2-D Range Data

Three-dimensional LiDAR can provide dense point clouds and, thus, has been predominantly applied in the field

of autonomous driving to deliver reliable person detection results [18], [19], [20], [21]. However, the expensive purchase and difficult data processing and storage of 3-D LiDAR make it challenging to apply in service robotics. In contrast, 2-D LiDAR at a relatively affordable price has become a more feasible deployment option. Nevertheless, the sparsity of 2-D range data creates complications for human feature extraction [22].

The person detection approaches based on 2-D range data could be mainly divided into two types. One is the traditional handcraft-based template matching technique. Predesigned heuristics, such as human leg [5], [6] and knee shape [7], numerical measurement [8] and clustering [9] features, and so on, are employed to extract feature sets from 2-D point clouds and then to achieve applications, including human detection, localization, or tracking through trained classifiers. Nonetheless, the complex social behaviors in real-world crowds can easily make these methods obscured and confused, even lead to human-machine conflicts.

The other one is data-driven methods that learn the optimal features of humans through massive annotated samples, showing tremendous advantages in detection performance and is selected as our primary algorithmic framework. The DROW detector [10], [11] is the first deep-learning-based model that locates persons from 2-D range data. A 1-D convolutional network is designed to extract spatial features from range measurement that correspond to humans. To better incorporate historical scan frames, DR-SPAAM [12] introduces a time-gated mechanism using dot-product attention. Given the difficulties in point cloud data annotation, a self-supervised algorithm based on pseudo-labels [14] is proposed, yielding performance comparable to supervised learning. However, the drawback of these methods is their narrow focus on a few aspects of temporal, spatial, and batched dimensions, resulting in insufficient fusion. To enhance detection performance, we propose the ODA module for omni-axes feature interaction.

B. Transformer in Point Cloud Analysis

The Transformer is considered a graph-based modeling structure in point cloud analysis, with its core self-attention mechanism modeling the relationships among different laser points in variable-length point cloud sequences, thereby capturing relevant features for the target objects. Previous studies have successfully applied the Transformer to point cloud processing tasks, demonstrating excellent flexibility and potential. For 3-D representation learning, remarkable progress has been made by models, such as MSPCT [23] and PointFormer [24], with notable utility in object scanning and indoor applications. In the field of 3-D object detection, VoTr [25] employs sparse voxels to capture long-range contextual information while maintaining computational efficiency similar to convolutional operations. SWFormer [26] realizes multiscale feature fusion through window shifting operations. In autonomous driving, Transformer-based models have demonstrated competitive performance in end-to-end environmental perception. Li3DeTr [27] leverages cross-attention modules to associate 3-D LiDAR global features with object predictions. Besides, DeepFusion [28] and TransFusion [29] employ a solution that

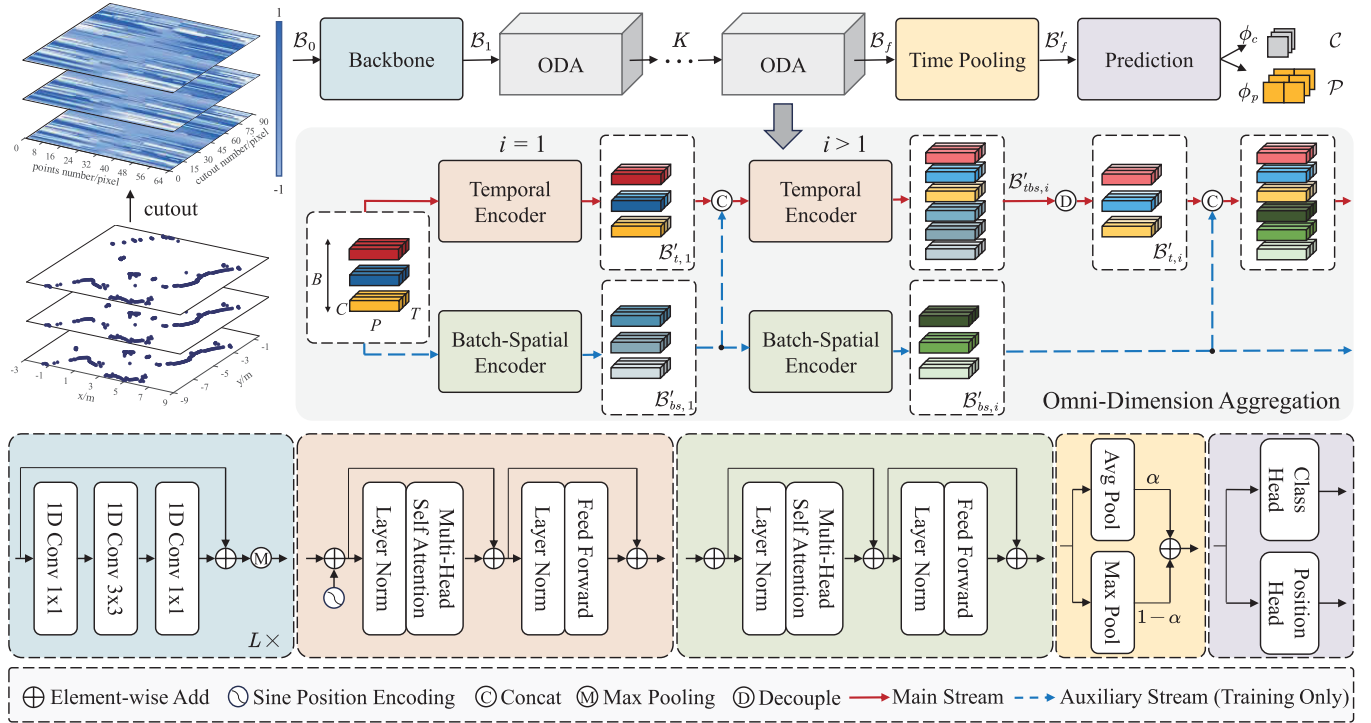


Fig. 2. Overall architecture of Li2Former. First, the preprocessed cutouts are fed into the backbone based on 1-D convolutional network. Next, the cascading ODA blocks further facilitate the information interaction across temporal, spatial, and batched dimensions, while the self-attention mechanism enhances the capability of global features modeling. Each ODA module consists of a main aggregation stream and a training only auxiliary feature stream. The data from the auxiliary stream are injected into the main stream to accomplish feature fusion. The time pooling module adaptively projects historical temporal features onto the current frame. Finally, classification and regression heads are applied to pedestrian detection.

fuses LiDAR and camera sensors for object queries in 3-D point clouds.

Inspired by the effectiveness of Transformers in 3-D point cloud analysis, we migrate the Transformer architecture to 2-D range data processing. Considering that the motion of pedestrians depends on both historical sequences and the influence of others in space, human localization models need to simultaneously consider the interframe temporal features and the intraframe spatial features. Meanwhile, the potential constraints of similar or shared parts among samples are crucial for sparse range measurement augmentation. Therefore, temporal Transformers and batched-spatial Transformers are designed for omni-axes encoding, and multiple alignment operations are employed to generate improved feature representations.

III. APPROACH

A. Problem Formulation

The 2-D person localization task can be regarded as estimating all potential human positions from the current planar point cloud using historical frames. Assuming a 2-D LiDAR returns N points in angle sequence with indices $\mathcal{I} = \{1, 2, \dots, N\}$, the laser scan is denoted as $\mathcal{S} = \{s_i^t | (\rho_i^t, \theta_i^t), i \in \mathcal{I}\} (t = 1, 2, \dots, T)$, where ρ_i^t and θ_i^t are the distance measurement and corresponding angle at time t , and T is the historical sequence length. The goal is to predict the pedestrian position and a confidence score $\mathcal{P} = \{p_i^{t_c} | (\rho_i^{t_c} + \Delta\rho_i, \theta_i^{t_c} + \Delta\theta_i, \psi_i^{t_c}), i \in \mathcal{I}\}$ at the current time $t_c = T + 1$. Here, $\Delta\rho_i$ and $\Delta\theta_i$

indicate the offset of the human center relative to the laser points $s_i^{t_c}$. If $\psi_i^{t_c}$ exceeds the threshold, person existence at that localization is suggested.

In light of the sparsity and inadequate perceptual resolution in point cloud scans $\mathcal{S}$ from 2-D LiDAR, exploiting information from different dimensions during the 2-D range measurement process helps uncover higher quality patterns. This motivated us to design a ODA module that integrates features across the temporal, spatial, and batched axes. As shown in Fig. 2, we first preprocess the raw scans into per-point perception windows using the cutout algorithm [10], [11], [12]. These cutouts are then fed into a backbone consisting of 1-D convolutional blocks to extract fine-grained structure. Subsequently, cascading multiple ODA blocks promotes the interaction of features across different dimensions and gradually fills semantic gaps. Simultaneously, self-attention mechanisms expand the ERF. An AFM-based loss function weights difficult and simple samples, accelerating model convergence. Finally, classification and localization heads are employed to encode features, enabling the separation of foreground pedestrians and background noise.

B. Planar Point Cloud Backbone

Since planar point clouds are essentially 1-D distance measurement signals, 1-D convolutional networks are chosen as the backbone to avoid unnecessary computational costs arising from dimension transformations when applying 2-D convolutional kernels [30]. The raw 2-D range data are not suitable for direct network input as the point cloud density

varies greatly with distance, namely, objects close to the LiDAR cover numerous points, while distant ones only contain a few laser beams. The imbalance between foreground and background data makes the model prone to overfitting on training scenes.

The cutout algorithm [10], [11], [12] assists in the pre-process of laser data. For 2-D LiDAR scan $\{s_i^t\}_{i=1}^N$ consisting of N points at time t , every N_0 points are selected as centers, and C local perceptual windows called *cutouts* are generated in Euclidean space, where the angular opening α_i^t of each cutout c_i^t is

$$\alpha_i^t = 2\arctan \frac{W}{2\rho_i^t} \quad (1)$$

where W denotes the hyperparameter for window width. All points inside the window will be clipped to a depth range of $\rho_i^t \pm H$. To ensure that each cutout contains an equivalent number of laser points, the measured data within α_i^t will be resampled to a fixed quantity M through linear interpolation. Each cutout will be mean-centered and normalized to $[-1, 1]$ based on ρ_i^t , meaning that the value of ρ_i^t does not affect the coverage range of the convolutional network's receptive field. This allows objects at different distances to exhibit roughly the same information characteristics. The 1-D convolutional backbone takes cutouts of continuous frames as input, extracting local features and mapping distance measurements into embedded vectors in a higher dimensional feature space

$$\mathcal{B}_1 = \text{Embedding}(\text{Conv}(\mathcal{B}_0)). \quad (2)$$

Herein, $\mathcal{B}_0 \in \mathbb{R}^{B \times T \times C \times M}$ with batch size B and time frames length T , and $\mathcal{B}_1 \in \mathbb{R}^{B \times T \times C \times P}$ with the dimensions of the embedded vectors P .

C. Omni-Dimension Aggregation

To tackle the limitation of insufficient information exchange caused by unidimension aggregation operators, cascaded ODA modules are developed to support feature fusion and alignment, resulting in superior feature representations.

1) *Two-Stream Training*: Inspired by [15], the ODA module is primarily divided into two data flows: 1) *main aggregation stream*, which receives extracted features from another data flow to enhance multidimensional information interaction and 2) *auxiliary feature stream*, which only conducts feature encoding within its own dimensions. The proposed Li2Former employs temporal encoders (T-Encoders) in fusion encoding and batched-spatial encoders (BS-Encoders) for intraframe spatial relationship extraction.

As shown in Fig. 2, the batch $\mathcal{B}_1$ from backbone output is duplicated as $\mathcal{B}_{t,1}$ and $\mathcal{B}_{bs,1}$ before entering the first ODA block, to be distributed to different streams, resulting in $\mathcal{B}'_{t,1}$ and $\mathcal{B}'_{bs,1}$. For the i th ($i > 1$) ODA block, the input of the main aggregation stream is concatenated with the output of the previous module, i.e., $\mathcal{B}_{tbs,i} = [\mathcal{B}'_{t,i-1}; \mathcal{B}'_{bs,i-1}]$, while the input of the auxiliary feature stream directly comes from the batched-spatial relationship of the previous blocks. Two-stream training pipeline can be seen as an internal implicit data augmentation, such as establishing potential connections for

similar or shared structure between samples, which is vital for sparse feature 2-D range measurements. By performing tensor concatenation, the interrelations among samples and point cloud spatial properties are injected into the main aggregation stream shared by the two data flows during the iteration process, contributing to the interaction and alignment of multidimensional information. With the main aggregation stream enhancing semantic similarity among omni-axes, Li2Former can remove the auxiliary feature stream during the test stage to expedite inference speed and reduce computational workload without causing significant performance degradation.

2) *Temporal Feature Extractor*: Dynamic changes from pedestrian state in the environment play an important role in human recognition and detection [31]. To extract the temporal dependencies between consecutive 2-D scans frames, a global T-Encoder is designed. For the first temporal encoding block, the computation process is as follows:

$$\mathcal{B}_{t,1}^{p'} = \text{MSA}_t(\text{LN}(\mathcal{B}_{t,1}^p)) + \mathcal{B}_{t,1}^p \quad (3a)$$

$$\mathcal{B}_{t,i}^{p'} = \text{FFN}(\text{LN}(\mathcal{B}_{t,i}^{p'})) + \mathcal{B}_{t,i}^{p'} \quad (3b)$$

where $\mathcal{B}^p = \mathcal{B} + \text{PE}(\mathcal{B})$ represents the cosine position encoding, as there is a clear temporal order between multiple point cloud frames. The multihead self-attention MSA_t is employed for time-series modeling with a sequence length of T . Residual connections and prelayer normalization (pre-LN) are widely used within modules to aid in model training [32]. A two-layer feedforward network (FFN) with ReLU activation is utilized to further map features to a higher dimensional feature space for temporal encoding. For the i th ($i > 1$) temporal encoding block, the computation process is as follows:

$$\mathcal{B}_{tbs,i}^{p'} = \text{MSA}_t(\text{LN}(\mathcal{B}_{tbs,i}^p)) + \mathcal{B}_{tbs,i}^p \quad (4a)$$

$$\mathcal{B}_{tbs,i}^{p'} = \text{FFN}(\text{LN}(\mathcal{B}_{tbs,i}^{p'})) + \mathcal{B}_{tbs,i}^{p'} \quad (4b)$$

$$\mathcal{B}'_{t(bs),i}, \mathcal{B}'_{(t)bs,i} = \text{decouple}(\mathcal{B}'_{tbs,i}). \quad (4c)$$

The distinction is that $\mathcal{B}'_{tbs,i}$ requires temporal-spatial decoupling to maintain its dimensionality, while preserving the temporal information $\mathcal{B}'_{t,i} = \mathcal{B}'_{t(bs),i}$. The batched-spatial feature $\mathcal{B}'_{(t)bs,i}$ will be replaced by data $\mathcal{B}'_{bs,i}$ from the auxiliary feature stream and concatenated as input $\mathcal{B}_{tbs,i+1} = [\mathcal{B}'_{t,i}; \mathcal{B}'_{bs,i}]$ into the T-Encoder of the next ODA block.

3) *Batched-Spatial Feature Extractor*: The BS-Encoder extracts both the spatial structure of point clouds and the similarity between different samples at the same time step. The computational process is as follows:

$$\mathcal{B}_{bs,i}^{p'} = \text{MSA}_{bs}(\text{LN}(\mathcal{B}_{bs,i}^p)) + \mathcal{B}_{bs,i}^p \quad (5a)$$

$$\mathcal{B}_{bs,i}^{p'} = \text{FFN}(\text{LN}(\mathcal{B}_{bs,i}^{p'})) + \mathcal{B}_{bs,i}^{p'} \quad (5b)$$

Multihead self-attention MSA_{bs} is utilized for batched-spatial sequence modeling with a sequence length of $B \cdot C$. Unlike T-Encoders, positional encoding is removed to achieve position-invariant learning of sample relations. Due to the presence of attention mechanism in the BS-Encoder, variable-length sequence inputs are acceptable; thus, Li2Former remains unaffected by different sensor configurations. From the perspective of optimization [15], the combination of batched and spatial dimensions alters the direction of gradient flow, allowing the spatial features of any given point clouds

to contribute to the loss computation of all other samples. Specifically, assume that the samples $\{\mathcal{B}^k\}_{k=1}^B$ with each mini-batch correspond to a network loss function value $\{\mathcal{L}^k\}_{k=1}^B$. In the collaborative learning process of BS-Encoder, we have

$$\frac{\partial \mathcal{L}^k}{\partial \mathcal{B}} = \frac{\partial \mathcal{L}^k}{\partial \mathcal{B}^k} + \sum_{j \neq k}^B \frac{\partial \mathcal{L}^k}{\partial \mathcal{B}^j}. \quad (6)$$

Indicating that the network implicitly incorporates gradients $\partial \mathcal{L}^k / \partial \mathcal{B}^j$ from virtual samples within mini-batch when processing data [33], which is a significant difference compared with the model that is optimized on a single batch. For data-imbalanced tasks, such as human localization based on 2-D range data (i.e., the ratio of target to background classes is uneven), virtual samples enhance data correlation and alleviate the feature sparsity. Moreover, to mitigate the bias between batch training and singleton inference, the BS-Encoder is set within the auxiliary feature stream, which can be removed during the testing phase.

4) *Time Pooling and Prediction*: With features alignment and fusion under the ODA blocks, composite semantic batch $\mathcal{B}_f \in \mathbb{R}^{B \times T \times C \times P}$ is generated as the output. Considering that only the current frame is involved in pedestrian prediction, an efficient way is required to compress the past $T-1$ frames to a lower dimensionality of $\mathcal{B}'_f \in \mathbb{R}^{B \times 1 \times C \times P}$. A dual-branch structured time pooling module is designed to adaptively project historical sequences onto the current scan. The maximum pooling branch focuses on local temporal features (e.g., clear person contours), while the average pooling branch emphasizes global ones (e.g., person movement trajectories). We introduce learnable parameters w_m and w_a to intelligently integrate information from different branch feature extractors

$$\mathcal{B}'_f = \frac{e^{w_m}}{e^{w_m} + e^{w_a}} \cdot \text{avg}(\mathcal{B}_f) + \frac{e^{w_a}}{e^{w_m} + e^{w_a}} \cdot \text{max}(\mathcal{B}_f). \quad (7)$$

Next, classification and localization heads with a two-layer MLP and ReLU activation are employed to generate pedestrian detection outcomes

$$\mathcal{C} = \phi_c(\mathcal{B}'_f; W_c) \quad (8a)$$

$$\mathcal{P} = \phi_p(\mathcal{B}'_f; W_p) \quad (8b)$$

where $\mathcal{P} \in \mathbb{R}^{B \times C \times 2}$ and $\mathcal{C} \in \mathbb{R}^{B \times C \times 1}$.

D. Adaptive Focusing Mechanism

In 2-D range data, the pedestrians to be detected can be classified into the following categories.

- 1) *Simple Samples*: Dense point clouds with clear contours.
- 2) *Challenging Samples*: Blurred features with occlusions.
- 3) *Invalid Samples*: Incorrectly labeled or excessively sparse.

A referenceless pursuit of the detector's fitting to the entire dataset can introduce negative gain from invalid samples, while failing to fully mine the intrinsic patterns of difficult ones, thus diminishing the model's convergence and generalization capabilities. Inspired by Tong et al. [34], an AFM based on Gaussian distribution is devised to alleviate issues of low-quality annotated samples.

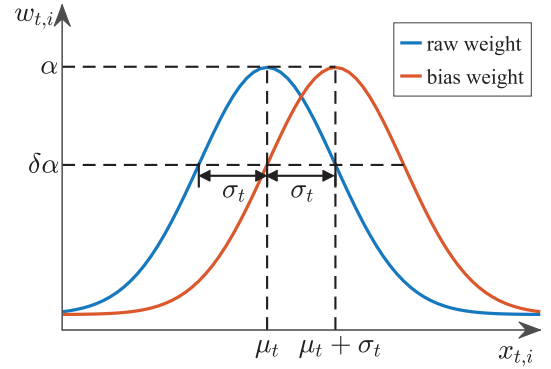


Fig. 3. Focusing weight curve demonstration. α is the focusing coefficient, and δ is the distances coefficient. During the training process, the focusing interval adaptively changes, enabling the loss function to guide the model to dynamically explore data patterns. The bias weight version provides higher attention to difficult samples, accelerating model convergence.

Let $\mathcal{L}_{\text{reg},t} \in \mathbb{R}^{BC \times 1}$ be the regression loss vector for the t th iteration, defined as the mean-squared error between predicted human positions and ground truth, while B and C are the batch size and number of cutouts. $\tilde{\mathcal{L}}_{\text{reg},t}^*$ denotes the historical average regression loss, defined as the exponential moving average of the current batch's mean loss $\tilde{\mathcal{L}}_{\text{reg},t}$

$$\tilde{\mathcal{L}}_{\text{reg},t}^* = \beta \cdot \tilde{\mathcal{L}}_{\text{reg},t-1}^* + (1 - \beta) \cdot \tilde{\mathcal{L}}_{\text{reg},t} \quad (9)$$

whereas β is the sliding factor, which can reduce the oscillations during early iterations due to poor model performance. $\tilde{\mathcal{L}}_{\text{reg},t}^*$ indicates that the training exhibits memorization, reflecting a more realistic model prediction. Subtracting $\tilde{\mathcal{L}}_{\text{reg},t}^*$ from $\mathcal{L}_{\text{reg},t}$ elementwise yields the measurement vector $\mathbf{x}_t$, whose component $x_{t,i}$ measures the relative improvement of the model's accuracy on sample i during the t th training iteration compared with its historical performance. The attention of $x_{t,i}$ is calculated using the nonmonotonic Gaussian function

$$w_{t,i} = \alpha \exp\left(\ln \delta \cdot \frac{(x_{t,i} - \mu_t)^2}{\sigma_t^2}\right). \quad (10)$$

Here, $\mu_t = \tilde{\mathcal{L}}_{\text{reg},t}$, while σ_t represents the standard deviation of the regression loss in $\mathcal{L}_{\text{reg},t}$. α is the focusing coefficient, and $\delta \in (0, 1)$ is the discount coefficient. Increasing the ratio α/δ helps enhance the discrimination in focusing on different types of samples, but it may introduce additional data imbalance. As shown in Fig. 3, the focusing curve illustrates that in the raw weights, it indicates an improvement in predictive performance when $x_{t,i} < \mu_t$, while opposite signifies a decline. According to the 3σ principle, AFM provides a more intelligent strategy for allocating gradient gains. It prioritizes the efficient gradients of challenging samples within $[\mu_t - \sigma_t, \mu_t + \sigma_t]$ over the suboptimal gradients of simple samples ($x_{t,i} < \mu_t - \sigma_t$) and the detrimental gradients of outliers ($x_{t,i} > \mu_t + \sigma_t$). Meanwhile, the focusing interval dynamically changes during the training process, enabling the loss function to guide the model to adaptively uncover data patterns.

The final loss function is the sum of the classification loss and the AFM-weighted regression loss based on the pedestrian

TABLE I

QUANTITATIVE COMPARISON RESULTS OF THE Li2Former AND THE SOTA BASELINE ALGORITHMS ON THE DROW AND JRDB DATASETS. WE USE BOLD TO INDICATE THE BEST RESULTS FOR EACH EVALUATION METRIC

Method	DROW [10]						JRDB [35]					
	AP _{0.3} ↑	AP _{0.5} ↑	Peak-F1 ↑	EER ↑	GFLOPs ↓	FPS ↑	AP _{0.3} ↑	AP _{0.5} ↑	Peak-F1 ↑	EER ↑	GFLOPs ↓	FPS ↑
ROS leg detector [36]	0.239	0.273	0.367	0.335	-	18.7	0.285	0.363	0.453	0.450	-	5.8
Arras (re-trained) [5]	0.413	0.437	0.476	0.460	-	24.5	0.376	0.445	0.487	0.484	-	7.6
Leigh (re-trained) [6]	0.480	0.552	0.563	0.557	-	25.4	0.337	0.521	0.553	0.552	-	12.0
DROW($T = 1$) [11]	0.677	0.685	0.658	0.652	11.37	83.7	0.696	0.732	0.682	0.676	13.80	78.3
DROW($T = 5$) [11]	0.679	0.698	0.656	0.651	27.09	72.0	0.723	0.765	0.714	0.709	32.87	58.5
+pseudo labels [14]	0.711	0.726	0.691	0.685	27.09	72.0	0.753	0.786	0.732	0.728	32.87	58.5
DR-SPAAM [12]	0.676	0.699	0.677	0.671	11.38	81.4	0.721	0.768	0.718	0.713	14.91	75.2
+pseudo labels [14]	0.698	0.711	0.710	0.708	11.38	81.4	0.744	0.787	0.728	0.728	14.91	75.2
Li2Former($T = 5$)	0.704	0.748	0.700	0.696	10.55	66.9	0.740	0.796	0.744	0.742	19.25	57.0
Li2Former-A($T = 5$)	0.725	0.764	0.720	0.718	10.55	66.9	0.749	0.814	0.750	0.746	19.25	57.0

label sets $\mathcal{Y}$ and position ground-truth sets $\mathcal{G}$

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \tilde{\mathcal{L}}_{\text{reg}} \quad (11)$$

where

$$\mathcal{L}_{\text{cls}} = -\frac{1}{|\mathcal{Y}|} \sum_{\substack{y_i \in \mathcal{Y} \\ c_i \in \mathcal{C}}} [y_i \log c_i + (1 - y_i) \log(1 - c_i)] \quad (12a)$$

$$\tilde{\mathcal{L}}_{\text{reg}} = \frac{1}{|\mathcal{G}|} \sum_{\substack{\mathbf{g}_i \in \mathcal{G} \\ \mathbf{p}_i \in \mathcal{P}}} w_i \|\mathbf{g}_i - \mathbf{p}_i\|_2. \quad (12b)$$

IV. EXPERIMENTS

In this section, we first describe two benchmark datasets used for model training and testing. Then, evaluation metrics are established, and implementation details are presented. Subsequently, qualitative and quantitative experiments are designed to compare the proposed model with SOTA methods, providing a comprehensive understanding. In-depth ablation studies are conducted to analyze the effectiveness of the devised core modules. Furthermore, we demonstrate the application of the model in person detection in real-world scenarios.

A. Experimental Setup

1) *Datasets*: DROW [10] is one of the few 2-D-LiDAR-based human detection datasets, where each frame contains 450 range measurements generated by the SICK S300. All hyperparameter selections are determined on the DROW validation set and utilized during the training process on other datasets. To provide a more thorough evaluation of the model's performance, we incorporated a more complex dataset JRDB [35], which comprises multiple indoor structured scenes with dense crowds and annotated pedestrian 3-D point cloud data. A 10-cm horizontal slice around human knee position is extracted to synthesize 2-D-LiDAR-like signals. Since 3-D point cloud data can be reasonably mapped to planar signals, Li2Former is not influenced by different data types. Any pedestrian labels in the datasets with fewer than four laser points are considered “ignored” and excluded from the evaluation process.

2) *Evaluation Metrics*: To fully assess the model performance, we selected the following evaluation metrics based on prior works [10], [11], [12]: AP_d reflects the average precision of human recognition, with a detection considered positive only if the unmatched ground truth and predicted location are within the associated distance d . Peak-F1 and equal error rate (EER) are used to measure the combined performance of the model in recall and precision. In addition, the computational throughput and inference speed are quantified by floating-point operations (FLOPs) and frame per second (FPS).

3) *Implementation Details*: All models are implemented in Python using PyTorch. GeForce RTX 4090 GPU and AMD EPYC 7V13 CPU are used for training. Testing is done on GeForce RTX 2060 GPU and Intel Core i7-10875H CPU. During training, Adam is selected as the optimizer with a batch size of 16 for 40 epochs. The learning rate is initialized at 10^{-4} with an exponential decay (after each iteration) to 10^{-6} . To differentiate the impact of AFM, Li2Former and Li2Former-A are developed. Li2Former-A integrates focal weights, whereas Li2Former and other baseline models adopt binary cross-entropy classification loss and L_2 -norm regression loss.

B. Quantitative

Table I reports the quantitative results of Li2Former and baseline methods on two benchmark datasets. To keep the comparison meaningful, the same cutout parameters are used for DROW [11], DR-SPAAM [12], and our models. Pretrained and fine-tuned using pseudo-labels [14], the enhanced versions DROW^p and DR-SPAAM^p stand for the top-performing results based on the CNN. The outcomes of ROS leg detector [36] with official parameters as well as the retrained Arras [5] and Leigh [6] detectors are also included in Table I.

In terms of accuracy, the proposed approach achieves the highest 76.4% AP_{0.5} and 72.5% AP_{0.3} above the baseline on the DROW dataset. Specifically, Li2Former exhibits a remarkable improvement over DROW^p and DR-SPAAM^p, with 3.8% and 5.3% higher in AP_{0.5} and 1.4% and 2.7% higher in AP_{0.3}, respectively. On the JRDB dataset, Li2Former outperforms DROW^p and DR-SPAAM^p by 2.7% and 2.6% in AP_{0.5}. However, due to the absence of local induction bias and

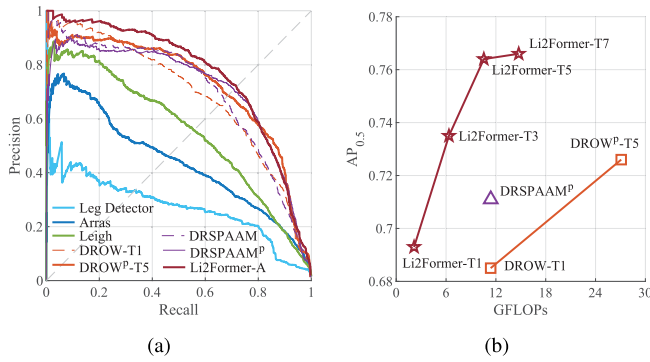


Fig. 4. (a) Precision–recall curves for several baselines and our Li2Former-A detector. (b) Trade-off between computation cost (GFLOPs) and model performance (AP_{0.5}). Superscript “p” denotes the usage of pseudo-labels, and “-T_i” represents the utilization of continuous i -frame scans. All models are evaluated on DROW test set with association distance 0.5 m, and our model outperforms the baseline within a reasonable computational complexity.

the loss of detailed contextual information caused by directly passing high-level features extracted by CNN backbone to Transformer, Li2Former shows a slight decrease of 0.4% in the finer AP_{0.3} compared with DROW^p, but maintains a 0.5% advantage over DR-SPAAM^p. Despite Li2Former’s global attention falls short in modeling details, its broader ERF presents a strength in pedestrian recall rate compared with CNN-based methods, leading to an increase in peak-F1 and EER. On the DROW dataset, peak-F1 and EER are 1% higher than the best result (DR-SPAAM^p), while on the JRDB dataset, peak-F1 and EER are 1.7% and 1.6% higher than the best result (DROW^p). Fig. 4(a) shows the precision–recall curves of all different models, observing that Li2Former performs well, except for a small region in the high recall range where DROW^p-T5 scores higher.

In terms of computational cost, traditional methods require enumerating human template matching results from multiple classifiers and performing weighted voting, which is unable to leverage GPU computation, further impeding its efficiency when processing large-scale point clouds. In contrast, end-to-end approaches demonstrate higher computational efficiency benefit from their parallel structures. The parallelism of attention calculation in the Transformer exceeds that of windowwise convolution, significantly reducing the complexity of Li2Former compared to DROW with the same sequential length scanning. Although DR-SPAAM’s RNN-based methodology alleviates the issue of increasing latency during inference by processing range data in a sequential manner across consecutive time frames, high-computational burden and extended training times still remain unavoidable. CNN-based models excel in decreasing model size through shared parameters in convolutional kernels. Conversely, the multihead self-attention modules and multilayer MLP in Li2Former produce an increase in parameter quantity, leading to slower inference speeds compared with DROW and DR-SPAAM. Fig. 4(b) compares the detection performance of Li2Former and the baseline across different scales. Notably, Li2Former-T1, which is equivalent to ignoring temporal encoding, displays lower accuracy, while Li2Former proves superior to the baseline in other settings, highlighting its effectiveness and scalability.

TABLE II

ABLATION STUDIES FOR ODA MODULE WITH DIFFERENT STRUCTURES ON DROW TEST SET. “M” REPRESENTS THE MAIN AGGREGATION STREAM, WHILE “A” STANDS FOR THE AUXILIARY FEATURE STREAM

Model	Structure	AP _{0.3}	AP _{0.5}	Peak-F1	EER	GFLOPs
1	Serial encoders	0.583	0.608	0.650	0.649	11.43
2	Parallel encoders	0.511	0.562	0.626	0.624	15.21
3	BS-Encoder(M)+T-Encoder(A)	0.621	0.662	0.698	0.688	10.67
4	BT-Encoder(M)+S-Encoder(A)	0.463	0.499	0.551	0.547	10.55
5	S-Encoder(M)+BT-Encoder(A)	0.471	0.509	0.565	0.561	10.67
6	T-Encoder(M)+BS-Encoder(A)	0.725	0.764	0.720	0.718	10.55

C. Qualitative

Fairly, we visualize the processing mechanism of different detectors on 2-D range data with a consistent setup. Fig. 5 illustrates that the testing indoor scenarios comprise of three categories: simple individuals (columns 1 and 2), crowds (columns 3 and 4), and humans affected by noise interference (column 5).

The repeated utilization of convolutional kernels enables a step-by-step enlargement of the ERF through signal propagation. Despite the benefits of the neighborhood property in extracting detailed features, convolutional operations limit efficient parallelization and long-range dependencies capturing. Template matching of human features using 1-D CNN enables the DROW detector to generate high responses at various pedestrian locations, but it encounters difficulties in suppressing the response of background points. As a result, the DROW detector can achieve high detection scores in simple scenes, but may produce multiple erroneous activations in complex environments due to confusion.

In contrast, Li2Former’s self-attention mechanism naturally models long-range data, providing a larger ERF to bring more spatial information, and improves the integrity of feature extraction, manifesting in the detector’s superiority in discriminating between background noise and foreground pedestrians. Especially, when noise exists in the scene, including static clutter or human-like barriers, Li2Former concentrates more on the details of object regions and ignores inaccurate correlations, leading to a greater identification capability and robustness.

D. Ablation Study

The core of Li2Former achieves progressive feature alignment and interaction in the temporal, spatial, and batched dimensions through cascaded ODA blocks, emphasizing the significance of selecting suitable ODA structures to elevate model performance. Moreover, the training of Li2Former with AFM retention and removal is applied to provide visual evidence of its effectiveness in enhancing learning efficiency.

1) *ODA Structure*: To validate the rationality of the ODA structure, we compare the detection performance of the model under different feature fusion strategies. As shown in Table II, the serial structure of Model-1 refers to using the output of T-Encoder as the input of BS-Encoder. The parallel structure of Model-2 involves directly fusing the outputs of T-Encoder and BS-Encoder using 1×1 convolutions. The remaining

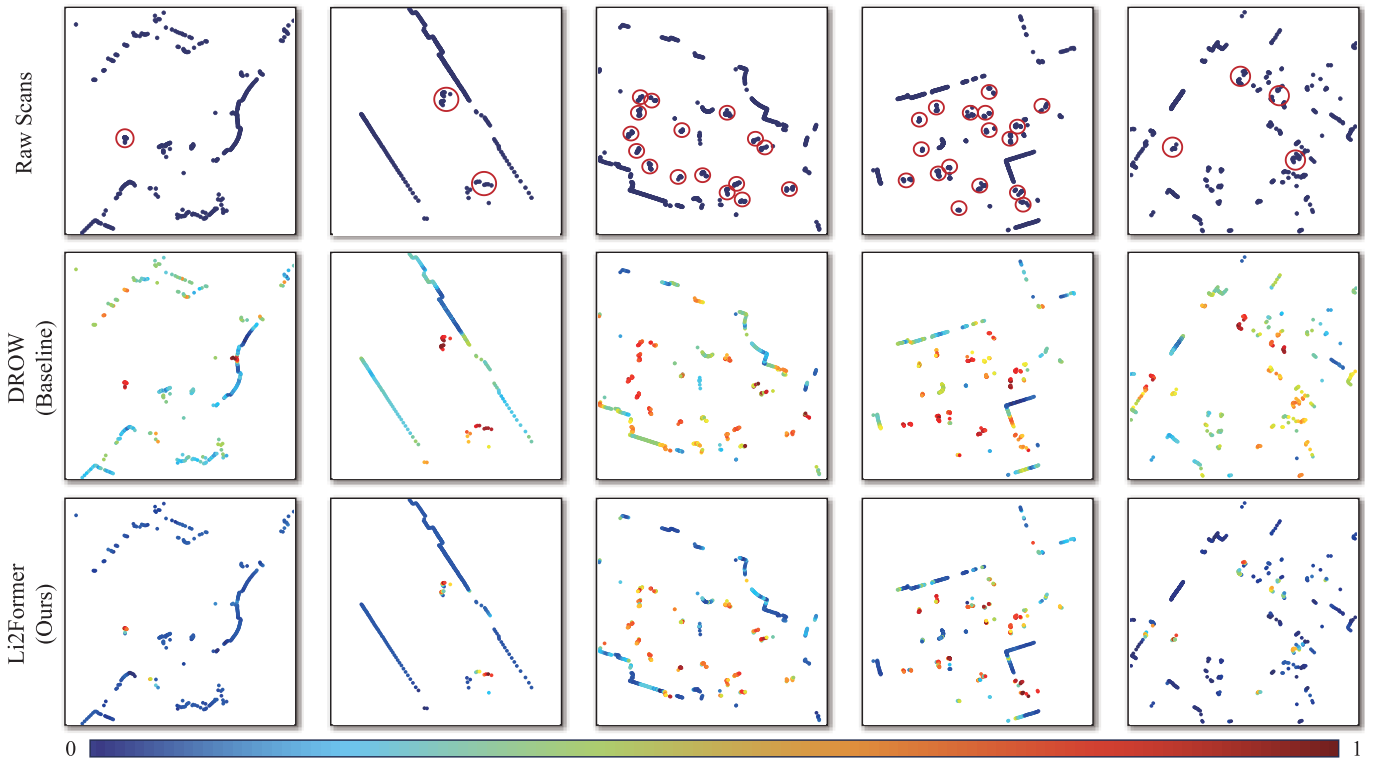


Fig. 5. Grad CAM [37] visualization of CNN-based baseline and our Li2Former in test point clouds. The first row contains the raw scans and annotated pedestrians, while the second and third rows depict the attention visualization of DROW and Li2Former, respectively, for different spatial laser scans. The different colors of the point cloud represent the varying degrees of attention paid by the detector, where colors tending toward red indicate stronger focus. Li2Former achieves better background and foreground separation performance compared with the baseline in different scenarios.

models employ two-stream training pipelines with different main aggregation streams and auxiliary feature streams.

Compared with the two-stream training pipeline, a Model-1's serial structure lacks internal data augmentation and only employs a naive fusion of temporal and spatial features. Model-2's parallel methodology overlooks the alignment of features across different dimensions, resulting in a decrease in model accuracy. Conversely, Model-3 and Model-6 showcase leadership in performance advancement. In Model-6, the adoption of the BS-Encoder in the auxiliary feature stream along with its integration into the T-Encoder of the main aggregation stream is more reasonable due to the inherent inconsistency between batch training and singleton inference, where employing BS-Encoder as the main aggregation stream can introduce bias during the testing stage. Both, based on two-stream pipeline, Model-4 and Model-5 exhibit a significant decline in performance. This can be attributed to the combination of the ordered temporal dimension and the unordered batched dimension, which disrupts the effectiveness of sequence learning and impacts the interpretability of the detection results.

The effect of various numbers of cascaded ODA blocks on model accuracy is conducted using the DROW validation set, as depicted in Table III. Compared with the test set, the marked discrepancy in detection scores is largely due to the greater prevalence of distant pedestrian samples in the validation set, where the sparsity of the 2-D range data makes model inference more challenging. The outcomes reveal that by increasing the ODA layer depth, the Li2Former model's

TABLE III
VALIDATION SET SCORES OF DROW FOR Li2Former DETECTOR
WITH DIFFERENT CASCADING LEVELS OF ODA

Number of Layers	AP _{0.3}	AP _{0.5}	Peak-F1	EER	GFLOPs
1	0.485	0.524	0.564	0.562	7.71
2	0.499	0.544	0.588	0.588	9.13
3	0.504	0.553	0.592	0.587	10.55
4	0.496	0.550	0.596	0.588	11.97
5	0.486	0.534	0.566	0.560	13.39
6	0.463	0.504	0.551	0.547	14.81

capacity is elevated, leading to better overall performance. Nevertheless, excessive ODA cascades can cause model overfitting, considering the data-hungry characteristics of the Transformer. To strike a balance between performance and computational complexity, we adopt a three-layer ODA module configuration in Li2Former.

2) *Two-Stream Training*: In order to prove feature fusion and alignment between the main aggregation stream and the auxiliary feature stream in the two-stream training framework, we extract the outputs of the T-Encoder and the BS-Encoder in each ODA module sequentially and calculate the semantic gaps, i.e., $d_i = \|\mathcal{B}'_{t,i} - \mathcal{B}'_{bs,i}\|_2^2$. Fig. 6(a) showcases the trend of d_i as the number of iterations progresses, disregarding the first ODA block at the input end as it solely involves samples duplication rather than feature fusion. Overall, the interstream feature gap in the ODA module diminishes gradually. The closer the ODA block (block 2) is to the output, the earlier

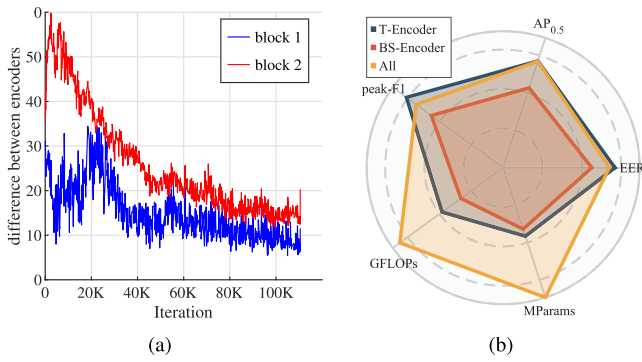


Fig. 6. Illustration of the effectiveness of the two-stream training pipeline. (a) Variation of feature differences between the encoder outputs located at different positions of ODA as the iteration number increases. (b) Comparison of computational complexity and accuracy between models that retain the main aggregation stream, auxiliary feature stream, and both.

TABLE IV
ABLATION STUDIES FOR TEMPORAL POOLING BLOCK WITH DIFFERENT POOLING METHODS ON DROW TEST SET

Method	AP _{0.3}	AP _{0.5}	Peak-F1	EER
Naive summation	0.680	0.732	0.691	0.689
Average pooling	0.703	0.751	0.695	0.688
1 × 1 convolution	0.669	0.727	0.686	0.682
Maximum pooling	0.657	0.710	0.665	0.663
Time pooling	0.725	0.764	0.720	0.718

d_i begins to decay, as gradient propagation becomes more effective toward the output end. The d_i ultimately converges to a constant value instead of decreasing to zero, because $B'_{t,i}$ encompasses temporal encoding information that is not present in $B'_{bs,i}$. Fig. 6(b) compares the computational cost and accuracy of the model with or without the auxiliary feature stream during the inference phase. Intuitively, retaining two data flows does not improve model accuracy but significantly increases model complexity and parameterization.

3) *Time Pooling*: Person prediction only takes into account the current laser frame, necessitating the design of a temporal fusion method to compress past $T - 1$ frames point clouds. Table IV examines the influence of different temporal aggregation policies on model performance. Naive summation, average pooling, and 1 × 1 convolution methods are equivalent to weighted averaging of range data over T time steps, with noticeably superior performance compared with the maximum pooling approach. Furthermore, the combination of maximum pooling and average pooling in the time pooling strategy yields the best results. This is because maximum pooling filters out time frames without detected pedestrians, allowing average pooling to focus more on essential person temporal information, thereby improving model generalization.

4) *AFM Guidance*: AFM is parameterized by the focusing coefficient α and the discount coefficient δ . Although the meaning of these two parameters is intuitively clear, with higher α/δ indicating greater discrimination by the model between challenging and simple samples, it remains valuable to select an appropriate hyperparameter combination. Table V illustrates the performance of Li2Former trained on different combinations of α and δ on the DROW validation set. The results indicate that increasing α/δ is generally benefi-

TABLE V
VALIDATION SET SCORES OF DROW FOR Li2Former DETECTOR WITH DIFFERENT AFM PARAMETERS

α	δ	AP _{0.3}	AP _{0.5}	Peak-F1	EER
1.5	0.50	0.499	0.541	0.585	0.580
1.5	0.33	0.504	0.553	0.592	0.587
1.5	0.25	0.494	0.541	0.589	0.585
1.0	0.50	0.495	0.537	0.589	0.587
1.0	0.33	0.495	0.538	0.583	0.582
1.0	0.25	0.492	0.533	0.582	0.581

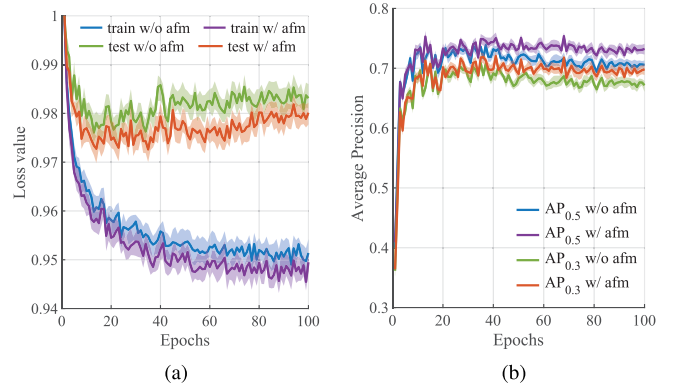


Fig. 7. Convergence and accuracy visualization of Li2Former with and without AFM on the DROW dataset. (a) Training and testing loss curves. (b) Average precision with association distance 0.3 and 0.5 m.

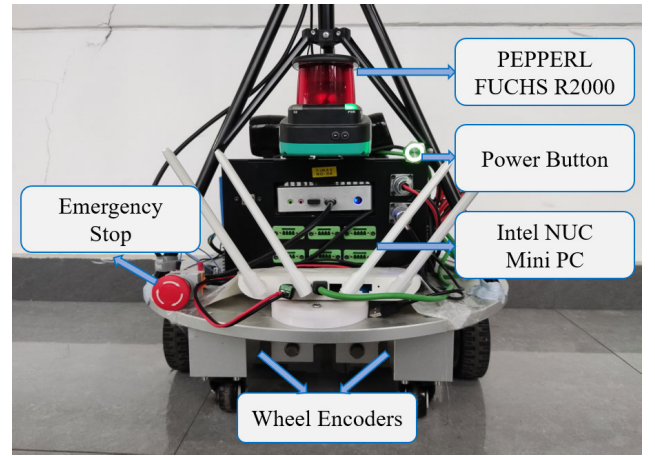
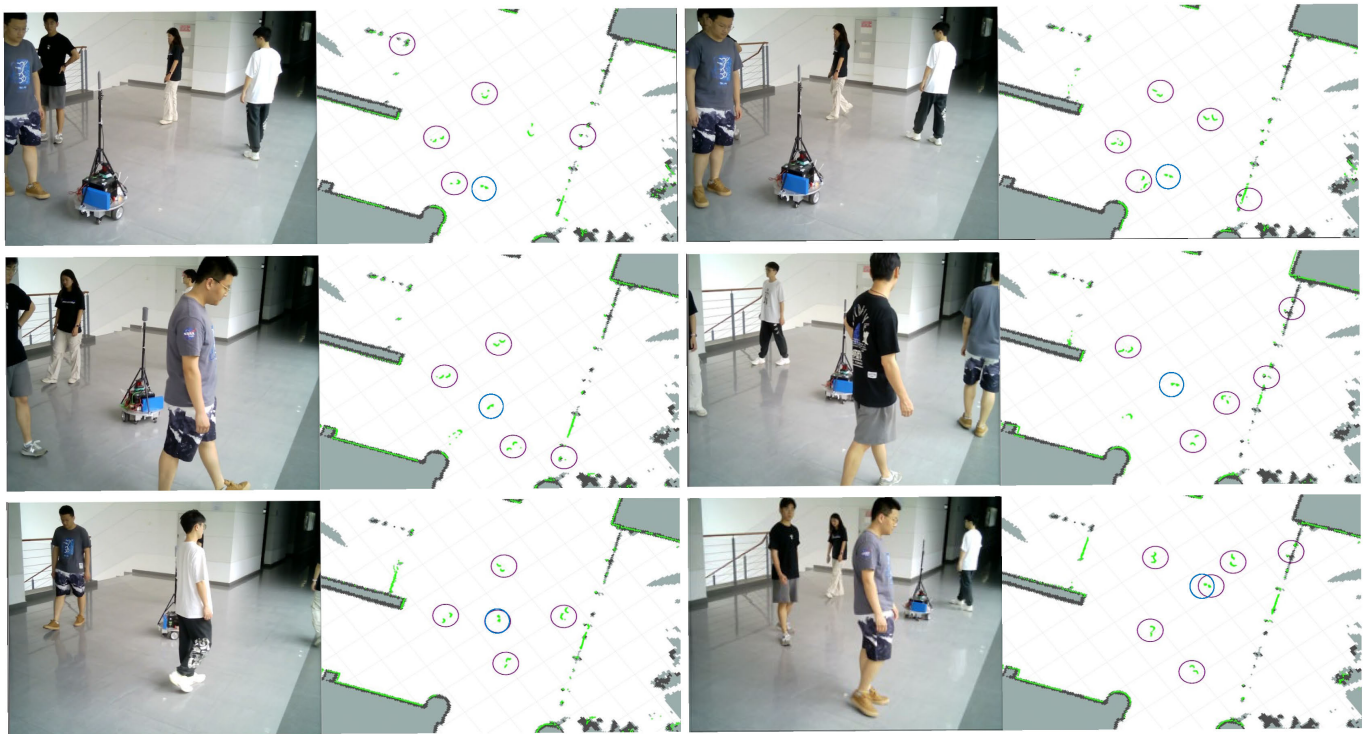


Fig. 8. Autonomous mobile robot for real-world experiments.

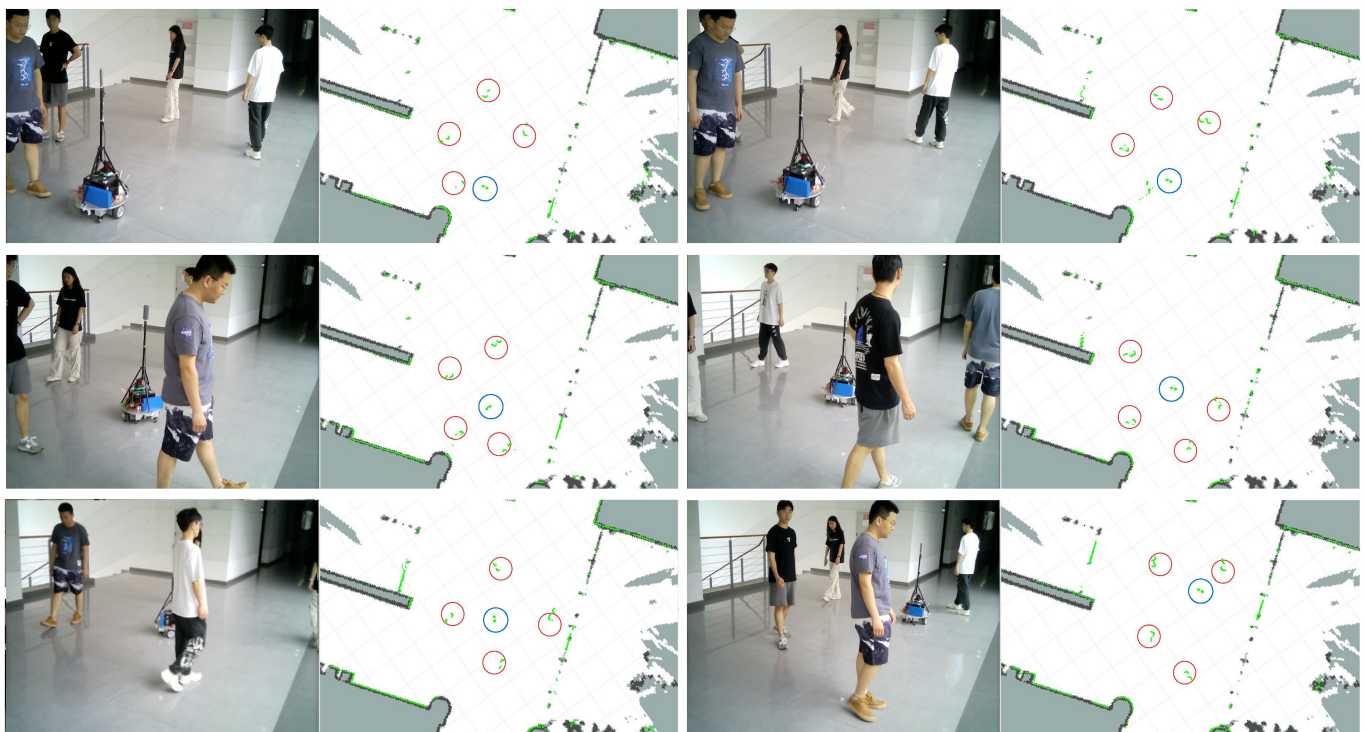
cial to improve model accuracy. However, further increments adversely affect performance, since the loss function becomes heavily focused on challenging samples, disregarding optimization for regular ones. On the other hand, higher α aids in guiding the network to learn patterns of difficult samples. Finally, we settle on $\alpha = 1.5$ as well as $\delta = 0.33$. Fig. 7 showcases the training and inference iterations of Li2Former with and without AFM, enabling a comprehensive evaluation of the effectiveness of AFM. Employing AFM facilitates a swifter reduction in model loss, converging to a lower level while preserving relatively higher accuracy.

E. Real World

Aside from the simulation above, we also deploy the trained detector into a differential wheeled robot to provide a thorough



(a)



(b)

Fig. 9. Real-world experiments validate the proposed method for pedestrians detection. Here, green points represent the raw laser scans, red/purple circles indicate detected human positions, and blue circle denotes the location of the robot. In contrast to the baseline model DROW, which tends to produce frequent false detections at humanoid laser positions, our approach is able to provide reliable results in most cases. (a) Detection results of DROW. (b) Detection results of our Li2Former.

validation in more realistic environments. The autonomous mobile robot is shown in Fig. 8. It is equipped with a 2-D LiDAR named PEPPERL FUCHS R2000, which has a

scanning range of 360° and a measurement radius of over 15 m. This kind of 2-D LiDAR offers accurate range data, high scan point density, and relative affordability, rendering it

TABLE VI
QUANTITATIVE COMPARISON OF THE Li2Former AND BASELINE
ALGORITHMS IN THE ACTUAL EXPERIMENTAL SYSTEM

Method	Memory(GB)	GFLOPs	FPS
DROW($T = 1$) [11]	1.45	14.81	50.7
DROW($T = 5$) [11]	1.76	35.27	35.8
DR-SPAAM [12]	1.95	15.05	49.3
Li2Former-A($T = 5$)	4.18	20.53	28.6

extensively adopted in indoor service robot applications. The usage of dual wheel encoders enables the provision of odometry feedback, and the emergency stop ensures the safety of human-machine interaction. The Intel NUC IPC with Intel i7-8700U processor and 16-GB memory is responsible for executing the low-level commands, including localization and motion control. In order to leverage GPU for inference acceleration to meet practical application demands, A laptop with a GeForce RTX 2060 GPU is used to perform sensor data processing and human detection.

Table VI presents the memory usage and detection latency of different methods. Due to the substantial parameter size resulting from the MSA and FFN in Li2Former architecture, the frame rate on actual device is lower compared with the baseline, which still basically meets practical requirements. Snapshots of real-world person detection using 2-D range data with the proposed method and traditional approach are shown in Fig. 9. During the experiment, all pedestrians move toward any destination at a normal walking speed. The results reveal that the Li2Former detector effectively identifies individual humans while avoiding false detections in the presence of background noise, achieving a lower false detection rate compared with DROW. However, restricted by the fact that the training sample only includes regular walkers, it can be observed that the model may fail when pedestrians are moving at high speeds (e.g., running). In addition, the loss of human features caused by special scanning angles results in high similarity between person and the environment, making it difficult for detector to perform foreground background separation. A hardware demonstration video can be found at <https://www.bilibili.com/video/BV1z64y1N7ta>.

V. CONCLUSION

In this work, we propose a novel Transformer-based omni-dimension aggregation module and a two-stream training framework to facilitate temporal, spatial, and batched feature alignment and interaction, addressing the limitations of unidimensional feature encoding and insufficient global perception in 2-D range-based person detection. Moreover, the introduction of an adaptive focusing mechanism to the loss function directs the model training to focus more on optimizing challenging samples, leading to improved model generalization and convergence speed. Ablation studies verify the effectiveness of the proposed ODA and AFM, and our approach achieves comparable results to the SOTA methods on two benchmark datasets. Furthermore, we conduct real-world validation of the model by deploying ROS-based packages on physical robots in actual environments.

However, while our approach demonstrating superior performance beyond the baseline, it is necessary to acknowledge the potential limitations of the proposed model. Primarily, the Transformer-based architecture leads to an increase in model parameters and a decrease in inference frame rate compared with CNN-based models, which might pose challenges in resource-limited scenarios. Besides, Li2Former has difficulty tracking pedestrians moving at high speeds as the rapid changes in human motion states make it difficult to capture the tight temporal correlations. Despite these considerations, Li2Former's performance improvements in 2-D point cloud-based pedestrian detection offer valuable advancements, and these shortcomings will be further improved in future work.

REFERENCES

- [1] N. Robinson, B. Tidd, D. Campbell, D. Kulić, and P. Corke, "Robotic vision for human-robot interaction and collaboration: A survey and systematic review," *ACM Trans. Hum.-Robot Interact.*, vol. 12, no. 1, pp. 1–66, Mar. 2023.
- [2] A. Bonci, P. D. C. Cheng, M. Indri, G. Nabissi, and F. Sibona, "Human-robot perception in industrial environments: A survey," *Sensors*, vol. 21, no. 5, p. 1571, Feb. 2021.
- [3] G. Chen et al., "Pseudo-image and sparse points: Vehicle detection with 2D LiDAR revisited by deep learning-based methods," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7699–7711, Dec. 2021.
- [4] F. Bu, T. Le, X. Du, R. Vasudevan, and M. Johnson-Roberson, "Pedestrian planar LiDAR pose (PPLP) network for oriented pedestrian detection based on planar LiDAR and monocular images," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1626–1633, Apr. 2020.
- [5] K. O. Arras, O. M. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2D range data," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3402–3407.
- [6] A. Leigh, J. Pineau, N. Olmedo, and H. Zhang, "Person tracking and following with 2D laser scanners," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 726–733.
- [7] T. Taipalus and J. Ahtainen, "Human detection and tracking with knee-high mobile 2D LiDAR," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Dec. 2011, pp. 1672–1677.
- [8] C. Weinrich, T. Wengelfeld, C. Schroeter, and H.-M. Gross, "People detection and distinction of their walking aids in 2D laser range data based on generic distance-invariant features," in *Proc. 23rd IEEE Int. Symp. Robot Human Interact. Commun.*, Aug. 2014, pp. 767–773.
- [9] X. Chen, J. Liu, J. Wu, C. Wang, and R. Song, "LoPF: An online LiDAR-only person-following framework," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.
- [10] L. Beyer, A. Hermans, and B. Leibe, "DROW: Real-time deep learning-based wheelchair detection in 2-D range data," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 585–592, Apr. 2017.
- [11] L. Beyer, A. Hermans, T. Linder, K. O. Arras, and B. Leibe, "Deep person detection in two-dimensional range data," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2726–2733, Jul. 2018.
- [12] D. Jia, A. Hermans, and B. Leibe, "DR-SPAAM: A spatial-attention and auto-regressive model for person detection in 2D range data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10270–10277.
- [13] M. Bouazizi, C. Ye, and T. Ohtsuki, "2-D LiDAR-based approach for activity identification and fall detection," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 10872–10890, Jul. 2022.
- [14] D. Jia, M. Steinweg, A. Hermans, and B. Leibe, "Self-supervised person detection in 2D range data using a calibrated camera," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13301–13307.
- [15] Z. Hou, B. Yu, and D. Tao, "BatchFormer: Learning to explore sample relationships for robust representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7256–7266.
- [16] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 367–376.

- [17] J. Kossen, N. Band, C. Lyle, A. N. Gomez, T. Rainforth, and Y. Gal, "Self-attention between datapoints: Going beyond individual input-output pairs in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 28742–28756.
- [18] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3D proposal generation and object detection from view aggregation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1–8.
- [19] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3D object detection from RGB-D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 918–927.
- [20] A. H. Lang, 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. (CVPR)*, Jun. 2019, pp. 12697–12705.
- [21] X. Zhang et al., "RI-fusion: 3D object detection using enhanced point features with range-image fusion for autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.
- [22] W. Wang, X. Chang, J. Yang, and G. Xu, "LiDAR-based dense pedestrian detection and tracking," *Appl. Sci.*, vol. 12, no. 4, p. 1799, Feb. 2022.
- [23] G. Zhou, D. Wang, Y. Yan, C. Liu, and Q. Chen, "6-D object pose estimation using multiscale point cloud transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.
- [24] X. Pan, Z. Xia, S. Song, L. E. Li, and G. Huang, "3D object detection with pointformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7463–7472.
- [25] J. Mao et al., "Voxel transformer for 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3164–3173.
- [26] P. Sun et al., "SWFormer: Sparse window transformer for 3D object detection in point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 426–442.
- [27] G. K. Erabati and H. Araujo, "Li3DeTr: A LiDAR based 3D detection transformer," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4239–4248.
- [28] Y. Li et al., "DeepFusion: LiDAR-camera deep fusion for multi-modal 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2022, pp. 17182–17191.
- [29] X. Bai et al., "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1090–1099.
- [30] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107398.
- [31] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "STGAT: Modeling spatial-temporal interactions for human trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6272–6281.
- [32] H. Xie, W. Song, and W. Kang, "MSBA-Net: Multiscale behavior analysis network for random hand gesture authentication," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.
- [33] Y.-Y. He, J. Wu, and X.-S. Wei, "Distilling virtual examples for long-tailed recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 235–244.
- [34] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IOU: Bounding box regression loss with dynamic focusing mechanism," 2023, *arXiv:2301.10051*.
- [35] R. Martín-Martín et al., "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.
- [36] C. Pantofaru. (2010). *Ros Leg Detector Package*. Accessed: Feb. 22, 2018. [Online]. Available: http://wiki.ros.org/leg_detector
- [37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.



Haodong Yang received the B.S. degree in control science and engineering from Tongji University, Shanghai, China, in 2022, where he is currently pursuing the M.S. degree in control science and engineering.

His research interests include light detection and ranging (LiDAR) simultaneous localization, navigation of mobile robots, and deep reinforcement learning.



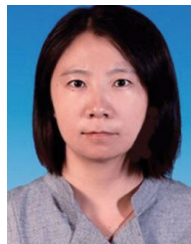
Yadong Yang received the B.S. degree from Tongji University, Shanghai, China, in 2022, where he is currently pursuing the M.S. degree with the College of Electronics and Information Engineering.

His main research interests include computer vision and image processing.



Chenpeng Yao received the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 2022.

He is currently a Post-Doctoral Researcher with the Postdoctoral Station of Mechanical Engineering, School of Automotive Studies, Tongji University, and joint educating at Tongji Artificial Intelligence (Suzhou) Research Institute. His research interests include the locomotion of biped robots, navigation of mobile robots, and reinforcement learning.



Chengju Liu received the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 2011.

From October 2011 to July 2012, she worked at the BEACON Centre, Michigan State University, East Lansing, MI, USA, as a Research Associate. From March 2011 to June 2013, she was a Post-Doctoral Researcher with the College of Electrical and Information Engineering, Tongji University, where she is currently a Professor. Her research interests include intelligent control, motion control of legged robots, and evolutionary computation.



Qijun Chen (Senior Member, IEEE) received the B.S. degree in automatic control from Huazhong University of Science and Technology, Wuhan, China, in 1987, the M.S. degree in control engineering from Xi'an Jiaotong University, Xi'an, China, in 1990, and the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 1999.

He was a Visiting Professor at the University of California at Berkeley, Berkeley, CA, USA, in 2008.

He is currently a Professor with the College of Electronic and Information Engineering, Tongji University. His current research interests include network-based control systems and robotics.