Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Enhancing Indoor Object Detection with xLSTM Attention-Driven YOLOv9 for Improved 2D-Driven 3D Object Detection

Yu He[1]*, Chengpeng Jin[2], Xuesong Zhang[1]

[1]School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian 116028, Liaoning, China
[2]School of Faculty of Land and Resources Engineering, Kunming University of Science and Technology, Kunming 650031, Yunnan, China

Corresponding author: Yu He, heyu3517@gmail.com

Abstract: Three-dimensional (3D) object detection is crucial for applications such as robotic control and autonomous driving. While high-precision sensors like LiDAR are expensive, RGB-D sensors (e.g., Kinect) offer a cost-effective alternative, especially for indoor environments. However, RGB-D sensors still face limitations in accuracy and depth perception. This paper proposes an enhanced method that integrates attention-driven YOLOv9 with xLSTM into the F-ConvNet framework. By improving the precision of 2D bounding boxes generated for 3D object detection, this method addresses issues in indoor environments with complex structures and occlusions. The proposed approach enhances detection accuracy and robustness by combining RGB images and depth data, offering improved indoor 3D object detection performance.

Keywords: Deep learning; Object detection; Attention mechanism

Online publication: February 24, 2025

1. Introduction

Currently, 3D object detection algorithms are widely applied in environmental perception systems, particularly in fields such as robotic control, intelligent surveillance, and autonomous driving. Recent 3D object detection algorithms typically rely on 3D sensors to capture spatial data [1-3]. Although traditional high-precision LiDAR and high-precision camera devices provide relatively accurate spatial information, their high costs and complex deployment requirements make them unsuitable for consumer-grade devices. In contrast, RGB-D sensors, exemplified by Kinect, have become an ideal choice for the consumer market due to their low cost and ease of deployment, making them well-suited to meet the demands of indoor environmental perception.

However, despite the significant cost advantage of RGB-D sensors over LiDAR, they still exhibit certain limitations in accuracy, especially in terms of precision and depth perception in complex indoor environments. Nonetheless, RGB-D sensors provide real-time RGB images and corresponding 3D depth information, making it possible to optimize 3D object detection algorithms using RGB images. In indoor object detection tasks, RGB images not only provide rich texture and color information but also help enhance detection robustness, particularly in low-light or occluded environments. Therefore, the integration of RGB images with 3D depth data holds significant research value and practical importance, particularly for the application of consumer-grade RGB-D sensors in indoor 3D object detection.

The F-ConvNet algorithm offers a promising approach by using 2D bounding boxes (2D BBoxes) generated by a 2D detector to slice the 3D space, thereby assisting in 3D object detection [4]. However, algorithms, such as F-ConvNet that rely on 2D detectors to generate 2D bounding boxes, have some shortcomings. Specifically, the generated 2D bounding boxes lack specificity and fail to fully account for the complexity and occlusions inherent in indoor environments. As a result, these methods struggle to precisely identify and localize indoor objects, particularly in narrow and complex spaces.

To address this issue, we propose an enhanced method that integrates the attention-driven YOLOv9 algorithm with xLSTM, incorporating it into F-ConvNet as the 2D detection module [5,6]. By introducing the attention mechanism of xLSTM, we can more accurately focus on and capture features relevant to 3D objects during the generation of 2D bounding boxes, especially in indoor environments. The temporal characteristics and long-short term memory capabilities of xLSTM enable it to better handle the dynamic changes and complex structures of indoor objects, thereby generating more precise and targeted 2D bounding boxes. This improvement significantly enhances the performance of F-ConvNet in indoor object detection, allowing the 3D object detection algorithm to more effectively combine RGB images and depth information, achieving higher detection accuracy and robustness.

In this paper, our main contributions can be summarized as follows:

(1) We integrate xLSTM into YOLOv9 to enhance its focus on indoor object detection, enabling it to generate more precise 2D bounding boxes (2D BBoxes). This improvement allows for better identification and localization of objects in indoor environments.

(2) We incorporate the xLSTM-enhanced YOLOv9 into the F-ConvNet framework, improving the accuracy of the 2D bounding boxes used for 3D object detection. This integration helps to better handle the complexities of indoor environments, providing more accurate and reliable detection results.

2. The proposed method

2.1. Vision-LSTM (ViL) + YOLOv9

Vision-LSTM (ViL) is a visual general-purpose network backbone constructed using xLSTM with residual blocks, as shown in Figure 1 [6]. The process is as follows: First, the input image is split into several patches, and then these patches are linearly projected. Each patch is augmented with a learnable positional information vector. Next, these patches are processed by alternating mLSTM blocks. For even-numbered blocks, the sequence is first flipped and then processed by the mLSTM layer. The processed sequence is normalized and finally passed through a linear projection for classification. The ViL encoder processes this sequence of patches, averaging the outputs of the first and last patch, and then the final classification result is output through a linear classification head.

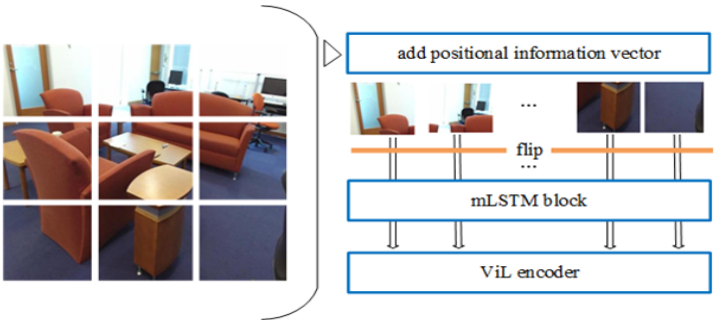

Figure 1. Vision-LSTM (ViL) architecture diagram

Vision-LSTM (ViL) utilizes mLSTM blocks to process image patches in a manner similar to time-series processing, introducing the concept of sequence processing into image tasks, akin to handling time-series data. We integrate Vision-LSTM (ViL) into YOLOv9 by introducing the EfficientLSTM module to enhance the temporal processing capability of image features. The ViLBlock and SequenceTraversal modules contained in EfficientLSTM are responsible for performing temporal processing of image features through mLSTM (multi-layer long short-term memory) networks. Specifically, the core component of ViL-mLSTM blocks is embedded into the feature extraction part of YOLOv9, helping the network capture dynamic changes and spatial relationships within the image.

2.2. xLSTM_YOLOv9 + F-ConvNet

In the process of integrating xLSTM_YOLOv9 into F-ConvNet, we begin by separating the RGB-D data into RGB and Depth components. The RGB data is then passed through xLSTM_YOLOv9, where the 2D detector generates 2D bounding boxes (2DBBoxes). These 2DBBoxes are subsequently mapped into the frustum generation structure of F-ConvNet to assist in the creation of the frustum. The entire workflow is illustrated in **Figure 2**.

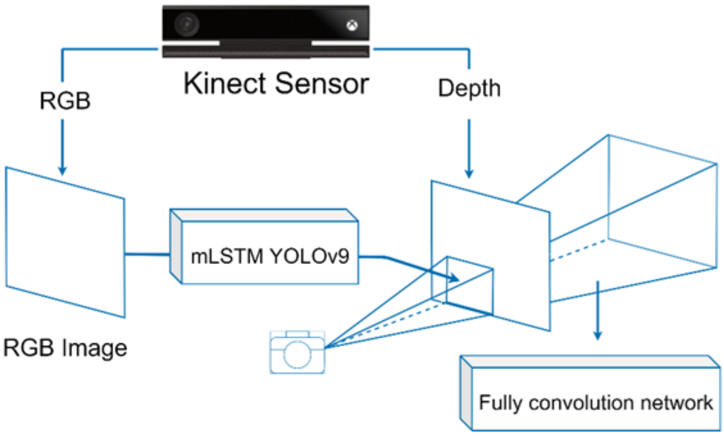

Figure 2. Integrating xLSTM_YOLOv9 into F-ConvNet

This approach leverages the strengths of both 2D and 3D detection, enabling more accurate object localization. By utilizing the RGB data for 2D bounding box generation through xLSTM_YOLOv9, we can refine the frustum generation process, ensuring it focuses on the most relevant areas of the scene. This integration not only improves detection accuracy in complex environments but also enhances the robustness of the system by efficiently combining the spatial and depth information. As a result, this method optimizes the handling of occlusions and varying object scales, making it especially effective in real-world 3D object detection tasks.

3. The proposed method

We integrated the xLSTM structure into YOLOv9 to enhance the performance of indoor object detection algorithms. To validate our proposed approach, we utilized the SUN RGB-D dataset. In this dataset, we selected 5050 scenes for validation and 5285 scenes for training. For the 2D experiments, we used the 2D data provided by the SUN RGB-D dataset [7]. The experimental results are presented in **Table 1**.

From **Table 1**, it is evident that after incorporating the xLSTM structure, the performance of the YOLOv9 algorithm in indoor object detection has been significantly improved, especially when dealing with objects that are prone to occlusion. The introduction of xLSTM allows the model to better capture temporal information, thereby enhancing its ability to recognize objects in dynamic scenes and complex environments. Compared to the traditional YOLOv9, the model with xLSTM exhibits stronger robustness in occluded object detection, small object recognition, and background noise suppression. This improvement is particularly crucial for real-world applications of object detection in indoor environments, where in complex scenarios with multiple objects interwoven, the xLSTM structure can effectively alleviate the recognition difficulties posed by occlusions and challenging backgrounds.

Table 1. Shortcut keys for the template

Method	bed	table	sofa	chair	toilet	desk	dresser	night stand	bookshelf	bathtub	mAP0.25
YOLOv9	0.81	0.46	0.61	0.71	0.86	0.33	0.48	0.69	0.54	0.60	0.61
xLSTM YOLOv9	0.85	0.55	0.69	0.76	0.87	0.35	0.49	0.69	0.54	0.52	0.64

In **Table 2**, we compare several typical RGB-D-based algorithms and also compare them with the F-ConvNet baseline algorithm, validating the effectiveness of our proposed approach. The evaluation metric used in **Table 2** is mAP@0.25 proposed by Song *et al.* [7].

Next, we integrated xLSTM_YOLOv9 into F-ConvNet, and the results are shown in **Figure 3**. In **Figure 3**, the first row displays the 2D RGB images from the SUN RGB-D dataset, the second row shows the detection results of the RGB scene using xLSTM_YOLOv9, and the third row presents the F-ConvNet detection results in the 3D point cloud scene, generated using the detection results from xLSTM_YOLOv9 and the frustum.

Table 2. Comparison of methods on different object categories

Method	bathtub	bed	bookshelf	chair	desk	dresser	night stand	soft	table	toilet	mean
DSS [8]	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG [9]	58.26	63.67	31.80	62.17	45.19	15.47	27.36	51.02	51.29	70.07	47.63
2Ddriven3D [10]	43.45	64.48	31.40	48.27	27.93	25.92	41.92	50.39	37.02	80.40	45.12
PointFusion [11]	37.26	68.57	37.69	55.09	17.16	23.95	32.33	53.83	31.03	83.80	45.38
Ren *et al.*[12]	76.2	73.2	32.9	60.5	34.5	13.5	30.4	60.4	55.4	73.7	51.0
F-PointNet [13]	43.3	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	90.9	54.0
F-ConvNet [4]	61.32	83.19	36.46	64.4	29.67	35.1	58.42	66.61	53.34	86.99	57.55
Ours	63.95	84.69	32.74	77.84	24.63	34.0	61.4	66.32	50.65	88.59	58.32

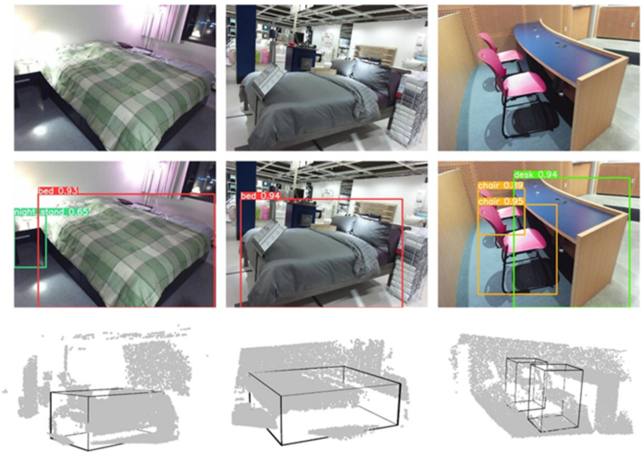

Figure 3. The detection result diagram of the F-ConvNet algorithm after integrating xLSTM_YOLOv9

Disclosure statement

The authors declare no conflict of interest.

Author contributions

Conceptualization: Yu He, Xuesong Zhang
Investigation: Yu He, Chengpeng Jin
Formal analysis and writing: Yu He, Chengpeng Jin

References

[1] Hu Y, Yang J, Chen L, et al., 2023, "Planning-Oriented Autonomous Driving". Proceedings of the IEEE/CVF Conference on CVPR, 2023: 17853–17862.

[2] Yang H, Zhang S, Huang D, et al., 2024, Unipad: A Universal Pre-Training Paradigm for Autonomous Driving. Proceedings of the IEEE/CVF Conference on CVPR, 2024: 15238–15250.

[3] Min C, Zhao D, Xiao L, et al., 2024, Driveworld: 4D Pre-Trained Scene Understanding via World Models for Autonomous Driving. Proceedings of the IEEE/CVF Conference on CVPR, 2024: 15522–15533.

[4] Wang Z, Jia K, 2019, Frustum ConvNet: Sliding Frustums to Aggregate Local Point-Wise Features for Amodal 3D Object Detection. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019: 1742–1749.

[5] Wang CY, Yeh IH, Liao HYM, 2024, YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. Computer–ECCV2024, 15089: 1–21.

[6] Beck M, Poppel K, Spanring M, et al., 2024, xLSTM: Extended Long Short-Term Memory. arXiv:2405.04517v1, 2024: 1–56.

[7] Song S, Lichtenberg SP, Xiao J, 2015, SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. Proceedings of the IEEE Conference on CVPR. Boston, Massachusetts, 567–576.

[8] Song S, Xiao J, 2016, Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images. Proceedings of the

IEEE Conference on CVPR, Las Vegas, Nevada, 808–816.

[9] Ren Z, Sudderth EB, 2016, Three-Dimensional Object Detection and Layout Prediction Using Clouds of Oriented Gradients. Proceedings of the IEEE Conference on CVPR, Las Vegas, Nevada, 1525–1533.

[10] Lahoud J, Ghanem B, 2017, 2D-Driven 3D Object Detection in RGB-D Images. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 4622–4630.

[11] Xu D, Anguelov D, Jain A, 2018, Pointfusion: Deep Sensor Fusion for 3D Bounding Box Estimation. Proceedings of the IEEE Conference on CVPR, Salt Lake City, Utah, 244–253.

[12] Ren Z, Sudderth EB, 2018, 3D Object Detection with Latent Support Surfaces. Proceedings of the IEEE Conference on CVPR, Salt Lake City, Utah, 937–946.

[13] Qi CR, Liu W, Wu C, et al., 2018, Frustum Pointnets for 3D Object Detection from RGB-D Data. Proceedings of the IEEE Conference on CVPR, Salt Lake City, Utah, 918–927.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.