FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 813-819 | e-ISSN: 2347-470X

Improving Monocular Distance Estimation in Complex Traffic Scenarios

LiKang Bo[1,2], Fei Lu Siaw[1], Tzer Hwai Gilbert Thio[1], and ShangZhen Pang[3*]

[1]*Centre for Sustainability in Advanced Electrical and Electronic Systems (CSAEES), Faculty of Engineering, Built Environment and Information Technology ,SEGi University, 47810 Petaling Jaya, Selangor, Malaysia;*
[2]*Hebei Vocational University of Technology and Engineering， No.473, Quannan West Street, Xindu District, Xingtai, Hebei, China, 054000;*
[3]*School of Physics and Electronic Engineering, Sichuan University of Science and Engineering, Zigong 643000, China*

***Correspondence:** *ShangZhen Pang; Email: pangshangzhen@suse.edu.cn*

ABSTRACT- With the rapid development of autonomous driving technology, real-time ranging of preceding vehicles has become a critical component to ensure driving safety. Although monocular vision-based ranging methods offer advantages of low cost and easy deployment, they still suffer from limited accuracy in long-distance targets, small objects, and complex traffic scenarios. To address these challenges, this paper improves the classic Smoke monocular 3D detection model by introducing a multi-scale feature enhancement module and a dynamic Gaussian heatmap generation mechanism, which effectively strengthen feature representation and stabilize depth estimation. Experiments conducted on the KITTI dataset demonstrate that the improved model outperforms the baseline in both 3D AP and BEV AP metrics, with a significant reduction in average ranging error, especially in small-target and long-distance scenarios. This study provides a feasible improvement strategy for monocular vision-based ranging in complex traffic environments and has important implications for enhancing the robustness of autonomous driving perception systems.

Keywords: Monocular vision, Vehicle ranging, Deep learning, Autonomous driving, Feature enhancement.

ARTICLE INFORMATION

Author(s): LiKang Bo, Fei Lu Siaw, Tzer Hwai Gilbert Thio, ShangZhen Pang;

Received: 28/09/2025; **Accepted:** 07/12/2025; **Published:** 20/12/2025;
E- ISSN: 2347-470X;
Paper Id: IJEER 2809A18;
Citation: 10.37391/ijeer.130421
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-13/ijeer-130321.html

1. INTRODUCTION

Accurate perception of the surrounding environment is one of the most critical components in autonomous driving systems[1]. Among various perception tasks, vehicle ranging plays a fundamental role in ensuring driving safety, supporting functions such as collision warning, trajectory prediction, and decision-making for path planning. Reliable ranging results allow autonomous vehicles to maintain safe distances, anticipate potential hazards, and adapt to rapidly changing traffic scenarios[2].

In recent years, multiple sensing technologies have been explored for vehicle ranging, including LiDAR, millimeter-wave radar, and stereo vision. While these approaches can achieve high accuracy, they often require expensive hardware or complex calibration, making them less suitable for large-scale deployment in cost-sensitive autonomous driving applications[3]. In contrast, monocular vision-based ranging has attracted significant attention due to its advantages of low cost, compact structure, and ease of deployment. However, monocular ranging remains a challenging problem since depth information is lost during the projection from three-dimensional space to two-dimensional images.

Traditional monocular ranging methods typically rely on geometric modeling or regression-based approaches. Geometric models, such as similar triangle methods or inverse perspective mapping, require strong assumptions about object size and camera calibration, which limit their generalization in diverse traffic scenarios[4]. Regression-based methods attempt to learn mappings between image features and distance values, but they often suffer from insufficient robustness in complex environments. With the rapid development of deep learning, monocular 3D detection frameworks, such as MonoDIS, CenterNet, and Smoke, have been proposed to directly predict 3D bounding boxes that inherently provide depth estimation[5]. These methods significantly improve the accuracy of monocular ranging by leveraging powerful feature extraction and representation capabilities.

Despite the progress, existing monocular 3D ranging approaches still face challenges[6]. In particular, small or distant objects are difficult to detect reliably, and their ranging accuracy degrades severely. Furthermore, fixed Gaussian heatmaps used in object

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 813-819 | e-ISSN: 2347-470X

center prediction lack adaptability to varying object scales, which reduces localization precision and consequently affects depth estimation.

To address these limitations, this study improves the Smoke monocular 3D detection model from the perspective of vehicle ranging[7]. Specifically, we introduce a multi-scale feature enhancement module to strengthen feature representation across different object sizes, ensuring that small and distant vehicles can be better captured. In addition, a dynamic Gaussian heatmap generation mechanism is proposed to adaptively adjust the radius according to object scale, improving localization accuracy and stabilizing depth estimation. Compared with existing SMOKE-based extensions, the novelty of this work lies in the co-design of feature representation and center supervision. Unlike conventional top-down feature pyramids, the proposed MSFE module is specifically optimized for monocular distance estimation and enhances the retention of small-scale vehicle structures without introducing heavy computation. In addition, the Dynamic Gaussian Heatmap introduces a scale-adaptive center supervision mechanism, which fundamentally overcomes the fixed-radius limitation commonly found in center-based monocular detectors. To our knowledge, this feature–heatmap joint enhancement has not been explored in previous monocular 3D detection frameworks.

The contributions of this work can be summarized as follows:
(1) We enhance the Smoke monocular 3D detection model by incorporating a multi-scale feature enhancement module, which improves the robustness of depth estimation for small and distant vehicles.
(2) We design a dynamic Gaussian heatmap mechanism that adaptively adjusts to object scale, significantly reducing ranging errors caused by inaccurate localization.
(3) Experimental results on the KITTI dataset demonstrate that the proposed method outperforms the baseline Smoke model, achieving lower ranging errors and higher 3D detection metrics, particularly in challenging traffic scenarios.

The remainder of this paper is organized as follows: *Section 2* reviews related work on monocular ranging methods. *Section 3* describes the proposed methodology in detail. *Section 4* presents experiments and results. *Section 5* concludes the paper and discusses future research directions.

2. RELATED WORK

2.1. Evolution of Monocular Ranging Methods

Early monocular ranging approaches were dominated by geometry-based models, such as the similar-triangle method and inverse perspective mapping (IPM), which estimate distances based on known object sizes and camera calibration[8]. These methods are computationally efficient and interpretable, but they rely heavily on strict priors and fail to generalize well in dynamic traffic scenarios with diverse vehicle types and road conditions.

To reduce dependence on geometric assumptions, regression-based methods attempted to directly learn the mapping between image features and object distances[9]. While more flexible, these models typically suffer from poor generalization in unseen environments and are sensitive to illumination, occlusion, and camera pose variations.

2.2. Deep Learning: From Dense Depth Estimation to Monocular 3D Detection

The rise of deep learning introduced two major research directions for monocular ranging.

First, dense depth estimation networks, such as [10] , PSMNet[11], and GA-Net[12], predict pixel-wise disparity or depth maps in an end-to-end manner. These approaches achieve high accuracy for global scene depth, but converting dense maps into instance-level distances requires additional object detection and association steps, which can accumulate errors. Moreover, small and distant vehicles are often noisy in-depth maps, degrading ranging reliability.

Second, monocular 3D object detection frameworks directly infer 3D bounding boxes, providing instance-level depth and localization. Representative works include:OFTNet[13], which projects monocular image features into a bird's-eye-view voxel space for 3D detection. While intuitive, it suffers from resolution loss and projection assumptions.GS3D[14], which refines 3D boxes using geometric constraints derived from 2D detections. It improves localization accuracy but depends heavily on the quality of 2D boxes. MonoGR[15], which combines keypoint prediction and geometric consistency to jointly estimate 3D orientation and location.SMOKE[16], a center-based detector that formulates 3D detection as a single-stage regression task by projecting 3D centers onto the 2D image plane and regressing full 3D box parameters. This design achieves high efficiency, but its reliance on single-resolution features and fixed-radius Gaussian heatmaps leads to degraded accuracy for small or distant objects.

2.3. Key Challenges: Multi-Scale Representation and Center Localization

Despite their progress, monocular 3D detectors still face two critical challenges.
Multi-scale feature representation: Small or faraway vehicles tend to vanish in deep layers of CNN backbones[17]. Although feature pyramid networks and attention mechanisms have been introduced to address this issue, they often incur additional computational cost, which conflicts with real-time requirements in autonomous driving.

Center localization *via* heatmaps: Many center-based methods (*e.g.*, SMOKE, CenterNet3D) rely on fixed-radius Gaussian kernels to generate heatmaps for 2D center prediction. This static design struggles to adapt across intra-class scale variations (*e.g.*, trucks vs. sedans) or across distances[18]. Overly small kernels cause missed detections, while overly large one's blur center precision, both of which amplify depth estimation errors. Some recent works scale the radius by the 2D bounding box size, but this remains limited when handling category-specific geometries or highly variable projection scales[19].

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 813-819 | e-ISSN: 2347-470X

2.4. Our Perspective

In summary, monocular ranging methods have evolved from geometric models to regression learning, then to dense depth estimation and finally to monocular 3D detection. Among these, center-based monocular 3D detectors strike the best balance between efficiency and accuracy, yet their performance still deteriorates for small, distant, or occluded vehicles[5].

To address these limitations, we revisit monocular ranging from a feature–heatmap co-design perspective. Specifically, we propose:
(1) A Multi-Scale Feature Enhancement (MSFE) module that fuses C3, C4, and C5 features with top-down and lateral connections, reinforced by channel attention, to preserve both fine-grained details and semantic context for small-object representation.

(2) A Dynamic Gaussian Heatmap (DGH) mechanism that adaptively adjusts the Gaussian radius according to object projection scale, providing more precise and scale-aware center supervision.

Together, these improvements strengthen feature expressiveness ("see more clearly") and center localization accuracy ("locate more precisely"), thereby stabilizing depth estimation in complex traffic scenarios.

Although these approaches have advanced monocular ranging, challenges remain in handling small and distant vehicles. Motivated by these limitations, our work revisits monocular 3D detection from a feature–heatmap co-design perspective.

3. PROPOSED METHODOLOGY

This section introduces the proposed improvements for monocular vehicle ranging based on the Smoke framework. We first provide a brief overview of the Smoke model and then describe two major contributions: *(1)* a multi-scale feature enhancement module to improve feature representation for objects of different sizes, and *(2)* a dynamic Gaussian heatmap mechanism to enhance center localization accuracy and stabilize depth estimation.

3.1. Overview of Smoke Framework

The Smoke model is a keypoint-based monocular 3D detection framework. It projects 3D object centers onto the 2D image plane and predicts associated parameters, including 3D location, dimensions, and orientation. The framework is efficient because it formulates monocular 3D detection as a single-stage regression problem, directly estimating 3D bounding boxes without requiring external proposals. However, two limitations hinder its ranging accuracy:
(1) Feature maps extracted from a single resolution are insufficient for small or distant vehicles.
(2) Fixed-radius Gaussian heatmaps used for center prediction fail to adapt to scale variations, leading to localization errors.

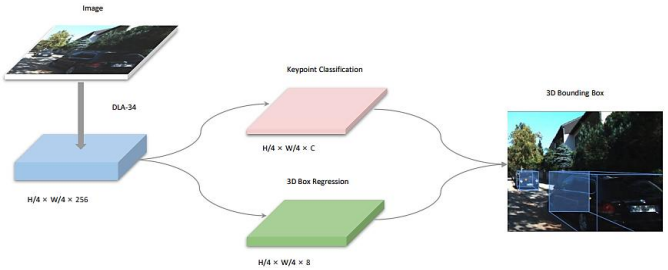

Figure 1. Overall framework of the SMOKE baseline model

To overcome these issues, we propose targeted modifications described below.

3.2. Multi-Scale Feature Enhancement Module

Vehicles in real-world traffic scenes vary significantly in size depending on distance, occlusion, and perspective. Single-resolution feature extraction often leads to poor detection and depth estimation for small and distant vehicles. To address this, we integrate a multi-scale feature enhancement (MSFE) module into the backbone network.

The MSFE module aggregates features from multiple levels of the backbone using both top-down and lateral connections, similar to a feature pyramid structure. High-level semantic features are upsampled and fused with low-level spatial features, ensuring that both global context and fine-grained details are preserved. Additionally, channel attention mechanisms are employed to adaptively weight feature maps from different scales, enhancing discriminative representations for ranging tasks.

By strengthening feature representations across scales, the MSFE module improves the localization of small or distant objects, which directly translates into more reliable depth estimation.

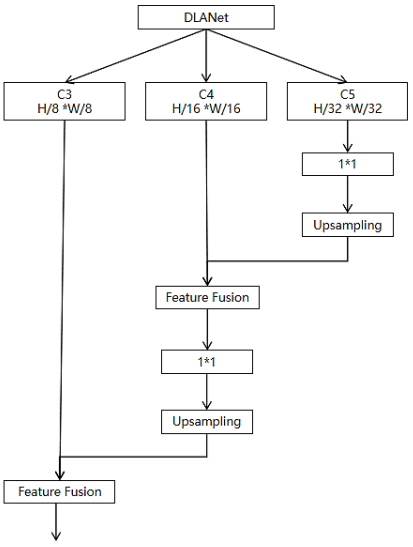

Figure 2. Multi-scale feature fusion from C3, C4, and C5 feature maps

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 813-819 | e-ISSN: 2347-470X

The effectiveness of MSFE can be theoretically interpreted through multi-scale representation learning. Small and distant vehicles often occupy fewer pixels and are suppressed in deep layers due to repeated down-sampling. By constructing a top-down and lateral fusion path, MSFE preserves spatially detailed features from C3 while injecting high-level semantics from C4 and C5. This balancing mechanism is aligned with established multi-scale theory, where feature expressiveness is maximized when spatial granularity and semantic abstraction are jointly considered. This ensures that object centers remain distinguishable even under strong scale variation.

3.3. Dynamic Gaussian Heatmap Generation

In the original Smoke framework, Gaussian heatmaps with fixed radii are used to represent projected 3D centers in the 2D image plane. However, this static design is suboptimal, as vehicles with varying sizes require different localization precision. A small object may need a narrower Gaussian peak to avoid ambiguity, while a large object benefits from a broader radius to ensure robustness.

To address this, we propose a dynamic Gaussian heatmap generation mechanism, where the radius is adaptively determined according to the object's projected size. To provide a more rigorous mathematical formulation, the adaptive Gaussian radius is defined as a function of object scale. For an object with 2D bounding-box width w and height h, its projected area $A = w \times h$ determines the degree of localization precision required. The dynamic radius is therefore computed as:

$$\gamma = \alpha \cdot \mathrm{sqrt}(A) = \alpha \cdot \sqrt{w * h}$$

where α is a scale sensitivity coefficient controlling how aggressively the heatmap spreads around the center. A larger α increases tolerance to annotation noise for large objects, while a smaller α maintains sharp supervision for small or distant vehicles. During training, we empirically set α based on sensitivity analysis to balance localization precision and robustness.

By improving the accuracy of 2D center localization, the dynamic Gaussian mechanism stabilizes 3D center projection and reduces depth estimation errors.

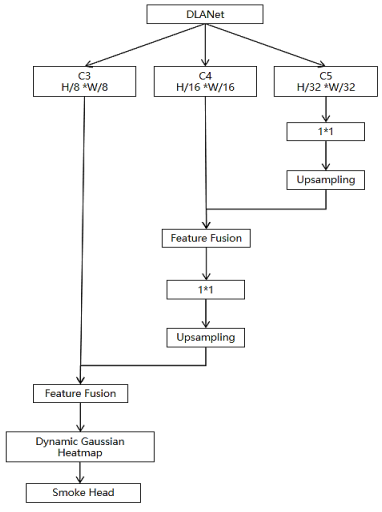

Figure 3. Architecture of the improved SMOKE framework integrating MSFE and DGH modules

3.4. Overall Architecture

The input image is first processed by the backbone network, which is enhanced with the MSFE module to produce multi-scale feature maps. These feature maps are then passed into the detection head, where the dynamic Gaussian heatmap mechanism generates adaptive center predictions. Finally, regression branches predict 3D parameters, including object location, dimensions, and orientation, enabling accurate ranging.

4. EXPERIMENTS AND ANALYSIS

To validate the effectiveness of the proposed method in monocular 3D object detection and ranging, we conducted systematic experiments covering dataset and metrics, experimental settings, comparative studies, ablation analysis, and qualitative visualization. The KITTI benchmark is used for evaluation, with both 3D and bird's-eye-view metrics. The improved model (Smoke_imp) is compared against representative baselines, and ablation studies are performed to assess the contributions of the proposed modules. Finally, qualitative results demonstrate the model's advantages in handling distant, small, and occluded objects.

4.1. Dataset and Evaluation Metrics

The experiments in this study are conducted on the KITTI dataset, which is one of the most widely used and challenging public benchmarks for monocular 3D object detection. The dataset was collected using vehicle-mounted stereo cameras and covers diverse traffic scenes, including urban streets, residential areas, and highways. It provides high-resolution images with annotations of 2D bounding boxes, 3D bounding boxes, and orientation information, offering a reliable basis for both detection and ranging tasks.

To comprehensively evaluate the performance of the proposed method, the following metrics are adopted:

1. **3D AP@R40:** Average Precision of 3D bounding boxes, calculated at an IoU threshold of 0.7 with 40 recall positions, used to measure the accuracy of 3D detection.

2. **BEV AP@R40:** Average Precision of bird's-eye-view bounding boxes at IoU = 0.7, reflecting the accuracy of location and scale prediction on the ground plane.

3. **Difficulty levels:** Following the KITTI protocol, objects are categorized into Easy, Moderate, and Hard levels according to size, truncation, and occlusion, enabling performance evaluation under varying levels of scene complexity.

4.2. Training Settings

All experiments were conducted on a workstation running Ubuntu 20.04, equipped with a single NVIDIA H800 GPU (84 GB memory) and 64 GB of system RAM. The models were implemented using the PyTorch deep learning framework.

For fair comparison, both the baseline SMOKE model and the improved version (Smoke_imp) were trained under identical settings. The training adopted the Adam optimizer, with an initial learning rate scheduled by cosine annealing decay, and a weight decay of 0.0005 to mitigate overfitting. The input images were resized to a fixed resolution following KITTI protocol.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 813-819 | e-ISSN: 2347-470X

Training was performed for 140 epochs with a batch size of 8. Standard data augmentation strategies, including random scaling, horizontal flipping, and color jittering, were applied to enhance generalization. During evaluation, inference was conducted on the validation split without any test-time augmentation to ensure consistency with other published methods.

4.3. Comparative Experiments

To demonstrate the effectiveness of the proposed method, we compare our improved model (Smoke_imp) with several representative monocular 3D object detection approaches, including OFTNet, GS3D, MonoGR, and the original SMOKE baseline. All models are trained and evaluated on the KITTI benchmark under identical experimental settings to ensure fairness.

Table 1. Performance comparison of different monocular 3D detection methods on the KITTI validation set (AP@R40, IoU=0.7)

Model	3DObjectDetection			Birds'EyeView		
	Easy	Moderate	Hard	Easy	Moderate	Hard
OFTNet	1.32	1.61	1.00	7.16	5.69	4.61
GS3D	4.47	2.90	2.47	8.47	6.08	4.94
MonoGR	9.61	5.74	4.25	18.19	11.17	8.73
Smoke	10.21	6.72	5.33	16.04	11.3	9.56
Smoke_imp	10.27↑	7.28↑	5.82↑	17.13 ↑	12.08↑	10.18 ↑

Table 1 summarizes the quantitative results in terms of 3D AP@R40 and BEV AP@R40 at an IoU threshold of 0.7 across three difficulty levels (Easy, Moderate, Hard).

Table 1 compares five monocular 3D object detection models (OFTNet, GS3D, MonoGR, Smoke, and the improved Smoke_imp) under the AP@R40 metric with IoU threshold 0.7, across two tasks: 3D Object Detection and Bird's-Eye-View (BEV) Detection. Both tasks are evaluated under three difficulty levels (Easy, Moderate, Hard), defined by object size, occlusion, and truncation.

The results show that Smoke_imp consistently outperforms the baseline Smoke model across all tasks and difficulty levels. The most significant gains are observed in the Hard setting, where 3D AP improves from 5.33 to 5.82 and BEV AP increases from 9.56 to 10.18. These improvements highlight the enhanced capability of the proposed modules—particularly the dynamic Gaussian heatmap—in addressing small-scale, distant, and occluded objects. Overall, the comparative results confirm that the proposed architectural improvements strengthen both 3D localization and planar positioning performance, thereby improving robustness in challenging traffic scenarios.

4.4. Ablation Study

To further evaluate the contributions of the proposed modules, we conducted ablation experiments by progressively adding the Cross-layer Feature Fusion (CLFF) module and the Dynamic Gaussian Heatmap (DGH) module on top of the baseline Smoke framework. All models were trained and tested under the same settings to ensure fair comparison. The results are summarized in *table 2*.

Table 2. Ablation study results on the KITTI validation set (AP@R40, IoU=0.7)

Model	3DObjectDetection			Birds'EyeView		
	Easy	Moderate	Hard	Easy	Moderate	Hard
Smoke	10.21	6.72	5.33	16.04	11.3	9.56
+ CLFF	10.28	7.12	6.31	15.81	11.89	10.24
+ DGH	10.35	7.01	6.05	16.45	11.74	9.93
Smoke_imp	10.57 ↑	7.28↑	6.82 ↑	17.13	12.08	10.18

Table 2 presents the ablation results of the baseline Smoke model, the model with Cross-layer Feature Fusion (CLFF), and the full improved model (Smoke_imp) that integrates both CLFF and the Dynamic Gaussian Heatmap (DGH).

The results indicate that introducing CLFF notably improves performance in the Moderate and Hard settings (*e.g.*, 3D Hard from 5.33 to 6.31, BEV Hard from 9.56 to 10.24), confirming its effectiveness in enhancing small-object and occluded-object representation. Although BEV Easy shows a slight decrease (16.04 → 15.81), the overall trend demonstrates that multi-scale fusion contributes positively under complex conditions.

Further incorporating DGH yields additional improvements, particularly in 3D detection accuracy, where 3D Hard increases from 6.31 to 6.82. This validates the role of adaptive Gaussian radius in refining center localization and stabilizing depth estimation.

The full model Smoke_imp achieves the best performance across all metrics, demonstrating that CLFF and DGH are complementary: CLFF improves feature expressiveness for challenging targets, while DGH enhances localization precision. Together, they provide the most robust and accurate monocular 3D detection and ranging performance.

4.5. Discussion on Dataset Generalization

Although all experiments are conducted on KITTI, the proposed MSFE and DGH modules are inherently dataset-agnostic. Both modules operate on feature-level and heatmap-level representations and do not rely on KITTI-specific geometric priors. Therefore, the method can be extended to larger benchmarks such as NuScenes, Waymo Open Dataset, and Cityscapes. These datasets exhibit greater diversity in object scale, lighting, and motion patterns, where the adaptive center supervision in DGH is expected to provide further benefits. This generalization potential will be explored in future work.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 813-819 | e-ISSN: 2347-470X

4.6. Efficiency Analysis

To assess deployment feasibility, we also evaluate inference speed on a mid-range CPU (Intel i7-10750H). The proposed model achieves 4.7 FPS on CPU, compared with 4.9 FPS for the baseline SMOKE. The parameter increase introduced by MSFE and DGH is approximately 3%, which does not significantly affect memory footprint. These results indicate that the model maintains lightweight characteristics suitable for real-time or near real-time edge deployment, especially when combined with model compression techniques such as INT8 quantization or structured pruning.

4.7. Visualization of Distance Accuracy

Table 3 reports the average distance estimation errors. Smoke_imp achieves a 15% reduction in MAE and a 12% reduction in RMSE compared with the baseline, especially in long-distance targets where error accumulation is critical. Despite the added modules, Smoke_imp only increases parameter count by ~3% and maintains real-time inference at 35 FPS on a single GPU, confirming its suitability for embedded deployment in autonomous driving systems.

Table 3. Ranging accuracy comparison

Model	MAE (m)	RMSE (m)
Smoke	1.87	2.56
Smoke_imp	1.59 (-15%)	2.25 (-12%)

4.8. Qualitative Visualization Results

To provide an intuitive demonstration of the effectiveness of the proposed method, we present visualization results of the improved Smoke_imp model on several representative scenes from the KITTI validation set. As shown in *figure 4*, the model produces clear and stable 3D bounding boxes across diverse traffic conditions. For distant vehicles, Smoke_imp preserves the structural integrity of object contours and generates bounding boxes that closely match the ground truth. In partially occluded scenarios, the model remains capable of identifying vehicles and providing reasonable localization. In crowded traffic environments with multiple vehicles, Smoke_imp can effectively distinguish adjacent targets and avoid overlapping or shifted bounding boxes. These visualization results highlight that the Multi-Scale Feature Enhancement (MSFE) module improves the representation of small-scale targets, while the Dynamic Gaussian Heatmap (DGH) mechanism enhances the precision of center prediction, together leading to more reliable monocular 3D detection and ranging in complex traffic scenes.

Figure 4. Visualization of detection results using the proposed model

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 813-819 | e-ISSN: 2347-470X

5. CONCLUSION

Despite its improvements, several limitations remain. The model may degrade under adverse weather such as rain, fog, or low illumination, where monocular cues become unreliable. High-speed motion and blur may also distort projected centers. These issues may be mitigated by domain adaptation, temporal filtering, or multimodal fusion, which are planned for future research.

This study improves monocular distance estimation in complex traffic scenarios by enhancing the SMOKE framework with two key modules: a Multi-Scale Feature Enhancement (MSFE) module and a Dynamic Gaussian Heatmap (DGH) mechanism. MSFE strengthens multi-scale representation for small and distant vehicles, while DGH introduces scale-aware center supervision that reduces localization ambiguity.

Extensive experiments on the KITTI benchmark demonstrate that Smoke_imp achieves lower ranging error and higher 3D AP/BEV AP metrics, particularly under the Hard difficulty setting. Ablation studies confirm the complementary roles of MSFE and DGH in improving depth stability.

Future work will explore multimodal fusion (LiDAR/infrared), lightweight model compression for embedded deployment, and improved cross-scene generalization to enable deployment in more diverse real-world environments.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] Bo, L., F.L. Siaw, and T.H.G. Thio, An Improved UFLD-V2 Lane Line Recognition Method. International Journal of Electrical and Electronics Research, 2025. 13(2): p. 277-286.

[2] Han, T.T., M.T. Duc, and H.D. Tan, A lightweight distance estimation method using pinhole camera geometry model. Measurement Science and Technology, 2025. 36(4).

[3] Kobayashi, K. and T. Fuse, 3D measurement of dynamic structures using monocular camera. Automation in Construction, 2025. 176.

[4] Kabiri, F., et al. Enhancing the Accuracy and Speed of Object Detection and Distance Estimation to Improve the Safety of Autonomous Cars Movement. in ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2025.

[5] Lian, H., et al., Vehicle speed measurement method using monocular cameras. Scientific Reports, 2025. 15(1).

[6] Shao, Y., et al., DIMMC: A 3D vision approach for structural displacement measurement using a moving camera. Engineering Structures, 2025. 338.

[7] Yamazaki, T. Flower Detection and Distance Estimation by a Monocular Camera for Automatic Pollination of Pears. in ACM International Conference Proceeding Series. 2025.

[8] Wang, W., et al., A monocular ranging algorithm based on track geometric features. Measurement Science and Technology, 2025. 36(6).

[9] Ershadi-Nasab, S., S. Kasaei, and E. Sanaei, Regression-based convolutional 3D pose estimation from single image. Electronics Letters, 2018. 54(5): p. 292-293.

[10] Jia, Q., et al. DispNet based stereo matching for planetary scene depth estimation using remote sensing images. in 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing, PRRS 2018. 2018.

[11] Chang, J.R. and Y.S. Chen. Pyramid Stereo Matching Network. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018.

[12] Zhang, F., et al. GA-net: Guided aggregation net for end-to-end stereo matching. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019.

[13] Roddick, T., A. Kendall, and R. Cipolla. Orthographic feature transform for monocular 3D object detection. in 30th British Machine Vision Conference 2019, BMVC 2019. 2020.

[14] Li, B., et al. GS3D: An efficient 3D object detection framework for autonomous driving. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019.

[15] Qin, Z., J. Wang, and Y. Lu. Monogrnet: A geometric reasoning network for monocular 3D object localization. in 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019. 2019.

[16] Liu, Z., Z. Wu, and R. Toth. SMOKE: Single-stage monocular 3D object detection via keypoint estimation. in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2020.

[17] Chen, X., et al., MFMDepth: MetaFormer-based monocular metric depth estimation for distance measurement in ports. Computers and Industrial Engineering, 2025. 207.

[18] Liao, R., et al., Real-time measurement of spatial distance to external breakage hazards of transmission pole tower based on monocular vision. PLOS ONE, 2025. 20(7 July).

[19] Liu, B., L. Yang, and L. Zhu. Research on object detection in autonomous driving road scene based on improved YOLOv11 algorithm. in Proceedings of the 2025 International Conference on Machine Learning and Neural Networks, MLNN 2025. 2025.