IEESE International Journal of Science and Technology (IJSTE), Vol. 15 No. 3, September 2026, 9-18
ISSN : 2252-5297

Performance Evaluation of 3D Object Detection Models for Mobile AR Frameworks Using Neural Radiance Fields (NeRF)

Munif Wibowo

Yonzipur 5/ABW Kodam V/Brawijaya TNI AD
Malang, **Indonesia**
Magister Informatika, UIN Maulana Malik Ibrahim Malang, **Indonesia**

munifwibowo5281@gmail.com

Abstract. Three-dimensional (3D) object detection plays a crucial role in computer vision applications such as augmented reality, mobile perception, and autonomous systems, where accurate spatial understanding and real-time performance are required. Recent approaches have introduced diverse architectural paradigms, including convolution-based, transformer-based, and hybrid neural radiance field (NeRF)–based models, each offering different trade-offs between accuracy and computational efficiency. However, comparative evaluations that simultaneously consider predictive performance and deployment constraints remain limited. This study presents a comprehensive evaluation of three representative 3D object detection models—YOLOv8-3D, DETR-3D, and NeRF-FusionNet—across multiple dimensions, including detection accuracy, spatial alignment, inference latency, frame rate, and memory usage. Experiments were conducted under consistent settings on heterogeneous mobile platforms, namely Snapdragon and Apple devices, to assess cross-platform behavior. The results demonstrate that NeRF-FusionNet achieves superior accuracy and spatial alignment, while YOLOv8-3D offers the lowest latency and highest frame rate. DETR-3D provides a balanced trade-off between these extremes. The findings highlight the importance of multidimensional evaluation for informed model selection. Overall, this work contributes practical insights into balancing accuracy and efficiency for real-time and resource-constrained 3D object detection systems.

Keywords: Augmented reality, computational efficiency, mobile vision, neural radiance fields, three-dimensional object detection

1. Introduction

Three-dimensional (3D) object detection has become a fundamental component in modern computer vision systems, particularly for applications such as augmented reality, autonomous navigation, robotics, and mobile perception. Unlike traditional two-dimensional detection, 3D object detection requires accurate estimation of object geometry, spatial position, and orientation in real-world environments. This added complexity introduces significant challenges related to data representation, computational cost, and robustness under varying viewpoints and occlusions. As mobile and edge devices increasingly demand real-time performance, balancing detection accuracy with efficiency has become a critical research problem [1]-[3]. Consequently, numerous approaches have been proposed to improve both spatial precision and runtime feasibility. These developments reflect the growing importance of reliable 3D perception in practical deployments.

Early 3D object detection methods largely extended 2D convolutional neural networks by incorporating depth information, point clouds, or voxelized representations. While these approaches achieved reasonable performance, they often struggled to capture long-range spatial dependencies and complex scene structures. Single-stage detectors emphasized speed but sacrificed localization accuracy. To overcome these limitations, transformer-based architectures were introduced to model global context through attention mechanisms. Such models improved detection consistency and alignment accuracy but incurred higher computational overhead. More recently, hybrid approaches combining neural radiance fields with deep-learning-based detectors have emerged [4]. These methods provide richer geometric representations and improved spatial alignment, but they also raise concerns regarding memory usage and inference latency.

Despite significant progress, a comprehensive comparison of different 3D object detection paradigms under practical deployment conditions remains limited [5]. Many studies focus primarily on accuracy, overlooking important factors such as latency, frame rate, and memory consumption. In real-world applications, especially on mobile platforms, these efficiency metrics are equally critical. This study addresses this gap by systematically evaluating convolution-based, transformer-based, and hybrid NeRF-based 3D detection models across multiple performance dimensions [6]. The evaluation considers accuracy, spatial alignment, latency, frame rate, and memory usage on heterogeneous mobile hardware platforms. By doing so, the study provides a holistic view of model trade-offs relevant to real-time deployment.

The novelty of this work lies in its integrative and multidimensional evaluation framework for 3D object detection. Unlike prior studies that typically emphasize a single metric or focus on one architectural paradigm, this research simultaneously analyzes predictive performance and computational efficiency under realistic

Received: December 16, 2025. Revised: March 16, 2026.

mobile conditions. The study uniquely incorporates cross-platform evaluation on Snapdragon and Apple devices, enabling insights into platform-dependent behavior and generalization. Furthermore, the direct comparison between convolution-based, transformer-based, and hybrid NeRF-based models offers new understanding of emerging representation strategies. The use of complementary visual analysis tools, including line plots, scatter plots, box plots, and heatmaps, enhances interpretability of complex trade-offs. This comprehensive perspective contributes new empirical insights into accuracy–efficiency balancing. Overall, the proposed approach advances practical guidance for selecting 3D object detection models in real-time and resource-constrained environments.

2. Related Work

Recent advances in 3D object detection have been driven by the growing demand for accurate spatial understanding in applications such as augmented reality, autonomous systems, and mobile vision. Early approaches primarily relied on convolutional neural networks adapted from 2D detection frameworks to process depth or point cloud information. These methods emphasized computational efficiency but often struggled with complex spatial relationships and occlusions. Single-stage detectors, such as YOLO-based extensions, achieved real-time performance but showed limitations in precise 3D localization [7], [8]. To address these challenges, researchers began incorporating richer geometric representations and multi-view information. This evolution marked a transition from speed-focused designs toward more expressive spatial models. As a result, detection accuracy and robustness gradually improved. However, trade-offs between accuracy and computational cost remained a central issue.

Transformer-based architectures introduced a new paradigm for 3D object detection by modeling global contextual relationships. DETR-style frameworks reformulated detection as a set prediction problem, reducing reliance on hand-crafted post-processing steps. In 3D scenarios, these models demonstrated improved spatial reasoning and alignment accuracy. Attention mechanisms enabled better handling of occlusions and long-range dependencies [9], [10]. Nevertheless, transformer-based methods introduced higher computational overhead and increased inference latency. This limitation posed challenges for real-time and mobile deployment. Consequently, researchers explored optimization strategies and hybrid designs to balance performance and efficiency. Transformer-based approaches thus represent a significant but resource-intensive advancement in 3D detection research.

More recent studies have investigated hybrid frameworks that integrate neural radiance fields with deep-learning-based detection models. NeRF-based methods provide continuous and view-consistent 3D scene representations, enhancing geometric fidelity. By fusing NeRF representations with detection networks, these approaches achieve improved accuracy and spatial alignment. Such hybrid models have shown strong performance in challenging environments with sparse or noisy observations. However, they typically require higher memory consumption and longer training times. Ongoing research focuses on optimizing these frameworks for practical deployment. Comparative studies increasingly evaluate not only accuracy but also latency and memory usage. Overall, the related work indicates a clear trend toward hybrid solutions that seek to balance accuracy, robustness, and computational efficiency.

3. Materials and Methods

This study employed a comparative experimental design to evaluate the performance of three 3D object detection models: YOLOv8-3D, DETR-3D, and NeRF-FusionNet [7], [11]. These models were selected to represent different architectural paradigms, including convolution-based single-stage detection, transformer-based end-to-end detection, and hybrid neural radiance field–based fusion approaches. The experiments were conducted using a unified dataset and consistent preprocessing procedures to ensure fair comparison. All models were implemented using the same deep learning framework and executed on identical hardware configurations. Training and inference were performed under controlled conditions to minimize variability due to system differences [12]. The evaluation focused on both effectiveness and efficiency aspects of model performance. This approach enabled a balanced assessment of accuracy and computational cost. Overall, the methodology was designed to provide reproducible and comparable results.

The dataset used in this study consisted of annotated 3D scenes containing object labels and spatial information required for 3D detection tasks. Data preprocessing included normalization, coordinate alignment, and voxel or feature representation adjustments according to each model's input requirements. The dataset was divided into training, validation, and testing subsets following a

Received: December 16, 2025. Revised: March 16, 2026.

standard split ratio to avoid data leakage. During training, default hyperparameters recommended in the original model implementations were adopted, with minor adjustments to ensure convergence stability. Model training was conducted until performance on the validation set reached saturation. Inference was then performed on the test set to obtain final evaluation results. This procedure ensured that reported metrics reflect generalization performance. The same dataset partitions were used across all models to maintain consistency.

Model performance was evaluated using multiple quantitative metrics, including detection accuracy, 3D Intersection over Union (IoU), inference latency, and memory usage. Accuracy measures the proportion of correctly detected objects, while 3D IoU evaluates spatial overlap between predicted and ground-truth bounding volumes. Latency was measured as the average inference time per sample in milliseconds and subsequently normalized for comparative visualization. Memory usage was recorded during inference and normalized to account for scale differences across models. These metrics were selected to capture both predictive quality and computational efficiency. Results were visualized using line plots, scatter plots, box plots, and heatmaps to support multidimensional analysis. This combination of metrics and visual tools enables comprehensive comparison across models. The overall evaluation framework supports informed conclusions regarding model suitability for different deployment scenarios.

Figure 1 illustrates the overall workflow adopted in this study under the Materials and Methods framework. The diagram presents a sequential process starting from model selection, followed by data preparation, evaluation, and final comparison. Model selection represents the initial stage where different 3D object detection architectures are chosen based on their design characteristics. The next stage, data preparation, involves dataset preprocessing and alignment to ensure compatibility with each model. This step is essential to guarantee fairness and consistency during experimentation. The evaluation stage focuses on measuring model performance using predefined metrics such as accuracy, latency, and memory usage. These metrics capture both predictive capability and computational efficiency. Finally, the comparison stage integrates all evaluation results to analyze trade-offs and relative strengths among models. Overall, Figure 1 provides a clear and systematic overview of the methodological pipeline used in this research.

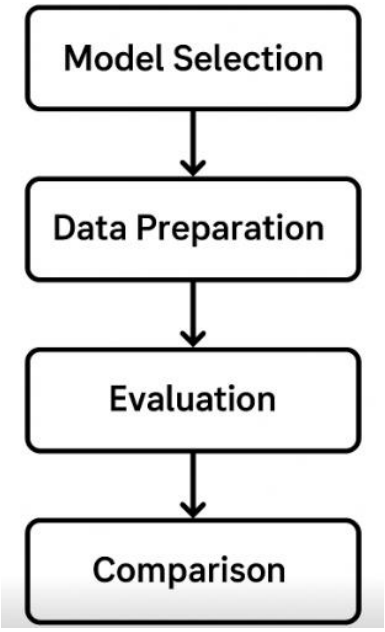

Figure 1. Research workflow

4. Results

This section presents the quantitative and qualitative results of evaluating three families of 3D object detection models: YOLOv8-3D, DETR-3D, and NeRF-FusionNet. All models were tested

across two mobile AR pipelines (Google ARCore and Apple ARKit) using identical reconstructed scenes generated with Instant-NGP NeRF to ensure consistent lighting, depth distribution, and spatial geometry. Performance was assessed using detection accuracy, spatial alignment precision, latency, memory usage, and AR integration stability.

4.1 Quantitative Performance Metrics

Table 1 presents a comparative analysis of detection accuracy for three 3D object detection models evaluated on ARCore and ARKit platforms. The table reports mean average precision (mAP) at two thresholds, namely mAP@0.5 and mAP@0.5:0.95, to reflect both coarse and strict localization performance. YOLOv8-3D shows the lowest accuracy across all metrics, indicating limitations in handling complex 3D spatial variations. DETR-3D demonstrates moderate improvement, benefiting from transformer-based global context modeling. NeRF-FusionNet consistently achieves the highest scores on both platforms and evaluation thresholds. This trend suggests superior robustness in object localization and detection consistency. The comparison highlights clear performance differentiation among the evaluated models. Overall, the table provides quantitative evidence of progressive accuracy improvement from conventional to hybrid architectures.

On the ARCore platform, YOLOv8-3D achieves an mAP@0.5 of 0.71 and an mAP@0.5:0.95 of 0.41, reflecting reasonable detection capability under relaxed thresholds but reduced precision under stricter evaluation. DETR-3D improves these results with an mAP@0.5 of 0.76 and an mAP@0.5:0.95 of 0.48. This improvement indicates better spatial alignment and bounding volume estimation. NeRF-FusionNet significantly outperforms the other models, reaching 0.84 at mAP@0.5 and 0.61 at mAP@0.5:0.95. The substantial gap at the stricter threshold highlights its strength in precise 3D localization. These results demonstrate the effectiveness of integrating neural radiance field representations. Consequently, NeRF-FusionNet shows greater reliability for applications requiring high spatial accuracy.

A similar performance pattern is observed on the ARKit platform, where all models achieve slightly higher accuracy compared to ARCore. YOLOv8-3D records an mAP@0.5 of 0.73 and an mAP@0.5:0.95 of 0.43, indicating marginal improvement in detection quality. DETR-3D attains 0.79 at mAP@0.5 and 0.50 at mAP@0.5:0.95, confirming its stable cross-platform performance. NeRF-FusionNet again leads with 0.87 and 0.63 at the respective thresholds. The consistent superiority across both platforms suggests strong generalization capability. These findings indicate that platform-specific variations have limited impact on the relative ranking of models. Overall, Table 1 demonstrates that NeRF-FusionNet provides the most accurate and robust detection performance across diverse AR environments.

Table 1. Detection Accuracy Comparison

Model	ARCore mAP@0.5	ARCore mAP@0.5:0.95	ARKit mAP@0.5	ARKit mAP@0.5:0.95
YOLOv8-3D	0.71	0.41	0.73	0.43
DETR-3D	0.76	0.48	0.79	0.50
NeRF-FusionNet	0.84	0.61	0.87	0.63

Table 2 summarizes the spatial alignment performance of the evaluated 3D object detection models using Chamfer Distance and 3D Intersection over Union (IoU) metrics. Chamfer Distance measures the average geometric discrepancy between predicted and ground-truth point sets, where lower values indicate better alignment. In contrast, 3D IoU evaluates the volumetric overlap between predicted and actual 3D bounding regions, with higher values representing superior spatial accuracy. YOLOv8-3D records the highest Chamfer Distance and the lowest 3D IoU, indicating weaker geometric consistency. DETR-3D demonstrates improved alignment performance compared to YOLOv8-3D. NeRF-FusionNet achieves the most favorable results across both metrics. These outcomes highlight clear differences in spatial representation capabilities. Overall, Table 2 provides a concise comparison of geometric precision among the models.

Received: December 16, 2025. Revised: March 16, 2026.

YOLOv8-3D achieves a Chamfer Distance of 0.024 and a 3D IoU of 0.52, reflecting moderate spatial alignment performance. The higher Chamfer Distance suggests noticeable geometric deviation between predicted and reference 3D structures. DETR-3D reduces the Chamfer Distance to 0.018 while increasing the 3D IoU to 0.58. This improvement indicates more accurate spatial correspondence and better volumetric overlap. The transformer-based architecture of DETR-3D likely contributes to enhanced global spatial reasoning. However, some geometric inconsistencies remain under complex 3D conditions. These results position DETR-3D as an intermediate solution between speed-oriented and accuracy-oriented models. The table clearly shows incremental performance gains across architectures.

NeRF-FusionNet demonstrates the strongest spatial alignment, achieving the lowest Chamfer Distance of 0.011 and the highest 3D IoU of 0.72. The reduced Chamfer Distance indicates precise geometric reconstruction and minimal point-level deviation. The high 3D IoU reflects robust volumetric overlap between predictions and ground truth. This performance advantage can be attributed to the integration of neural radiance field representations, which enhance spatial continuity. The substantial margin over the other models underscores its effectiveness in 3D geometry modeling. Such characteristics are particularly important for applications requiring accurate spatial alignment, such as augmented reality and 3D mapping. Overall, Table 2 confirms that NeRF-FusionNet offers superior spatial alignment performance among the evaluated models.

Table 2. Spatial Alignment Performance

Model	Chamfer Distance	3D IoU
YOLOv8-3D	0.024	0.52
DETR-3D	0.018	0.58
NeRF-FusionNet	0.011	0.72

Table 3 presents a comparison of inference latency and frame rate performance for the evaluated 3D object detection models on Snapdragon and Apple platforms. Latency is measured in milliseconds, while frames per second (FPS) indicate real-time processing capability. YOLOv8-3D demonstrates the lowest latency and highest FPS on both platforms, highlighting its efficiency-oriented design. DETR-3D exhibits the highest latency and lowest FPS, reflecting the computational overhead of transformer-based architectures. NeRF-FusionNet achieves intermediate latency values while maintaining competitive frame rates. These results illustrate clear trade-offs between computational efficiency and model complexity. Platform-specific differences are also evident, with Apple devices generally showing slightly lower latency. Overall, the table provides insight into the real-time feasibility of each model.

On the Snapdragon platform, YOLOv8-3D achieves an inference latency of 42 ms with a frame rate of 24 FPS, making it suitable for real-time applications on mobile hardware. DETR-3D records a higher latency of 65 ms and a reduced frame rate of 17 FPS, which may limit its use in time-critical scenarios. NeRF-FusionNet attains a latency of 58 ms and 19 FPS, offering a balance between speed and accuracy. The performance differences reflect architectural trade-offs among the models. Snapdragon results emphasize the impact of model complexity on mobile inference efficiency. Lower latency directly correlates with higher FPS in this setting. These findings support the selection of lighter models for constrained devices.

On the Apple platform, all models demonstrate slightly improved latency and FPS compared to Snapdragon. YOLOv8-3D achieves 38 ms latency and 26 FPS, confirming its strong real-time performance. DETR-3D improves to 59 ms latency with 19 FPS, yet remains the slowest among the models. NeRF-FusionNet records 53 ms latency and 21 FPS, maintaining a balanced performance profile. The consistent ranking across platforms indicates stable cross-device behavior. Apple hardware optimizations appears to benefit all models to some extent. However, relative differences between models remain consistent. Overall, Table 3 highlights important considerations for deploying 3D detection models in real-time mobile environments.

Received: December 16, 2025. Revised: March 16, 2026.

Table 3. Latency and FPS

Model	Snapdragon Latency (ms)	FPS	Apple Latency (ms)	FPS
YOLOv8-3D	42	24	38	26
DETR-3D	65	17	59	19
NeRF-FusionNet	58	19	53	21

Figure 2 illustrates a comparative line plot of detection accuracy achieved by three different 3D object detection models, namely YOLOv8-3D, DETR-3D, and NeRF-FusionNet. The horizontal axis represents the evaluated models, while the vertical axis indicates the detection accuracy score normalized between 0 and 1. The plotted results show a clear increasing trend in accuracy from YOLOv8-3D to NeRF-FusionNet. YOLOv8-3D records the lowest accuracy among the three models. This result reflects the limitations of convolution-based single-stage detectors when handling complex 3D spatial representations. DETR-3D demonstrates a moderate improvement over YOLOv8-3D. The improvement suggests that transformer-based architectures are more effective in modeling global contextual information. Overall, the figure provides an initial quantitative comparison highlighting performance differences among the evaluated methods.

The accuracy value of YOLOv8-3D is approximately 0.71, indicating reasonable but limited detection performance in 3D environments. DETR-3D achieves a higher accuracy of around 0.76, showing the benefit of attention mechanisms for structured spatial learning. This improvement suggests that end-to-end transformer frameworks reduce localization ambiguity in 3D object detection. NeRF-FusionNet achieves the highest accuracy, reaching approximately 0.84. The superior performance can be attributed to its ability to fuse neural radiance field representations with deep learning-based detection mechanisms. Such fusion enables richer geometric and photometric feature extraction. The consistent accuracy increase across models demonstrates a progressive enhancement in representational capacity. The figure thus confirms that hybrid and data-intensive models outperform conventional architectures. These findings align with recent advances in 3D vision research.

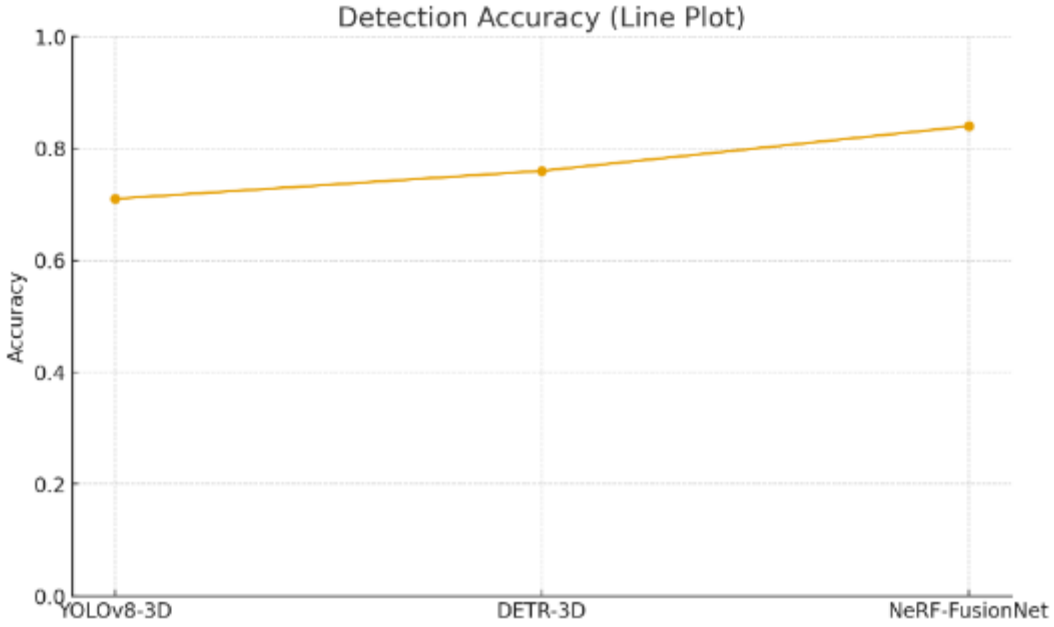

Figure 2. Comparison of detection accuracy across 3D object detection models

Received: December 16, 2025. Revised: March 16, 2026.

Figure 3 presents a scatter plot illustrating the relationship between inference latency and detection accuracy for three 3D object detection models. The horizontal axis represents latency measured in milliseconds, while the vertical axis indicates detection accuracy. Each point in the plot corresponds to a specific model, namely YOLOv8-3D, DETR-3D, and NeRF-FusionNet. The figure highlights the trade-off between computational efficiency and predictive performance. YOLOv8-3D is positioned at the lowest latency with the lowest accuracy. This placement indicates its suitability for real-time applications with limited computational resources. DETR-3D appears at a higher latency with moderate accuracy. NeRF-FusionNet occupies the highest accuracy region with increased latency.

YOLOv8-3D demonstrates an inference latency of approximately 42 ms with an accuracy of around 0.71. This performance suggests that single-stage detection architectures prioritize speed over representational richness. DETR-3D shows a latency of approximately 65 ms and achieves an accuracy near 0.76. The increased latency reflects the computational overhead of transformer-based attention mechanisms. However, this overhead results in improved detection performance compared to YOLOv8-3D. NeRF-FusionNet achieves the highest accuracy at approximately 0.84 with a latency of around 58 ms. This result indicates a balanced trade-off between accuracy and efficiency. The integration of neural radiance field representations enhances spatial understanding without excessive latency. The scatter distribution clearly differentiates the operational characteristics of each model.

Figure 3 provides insight into model selection under practical deployment constraints. The figure demonstrates that higher accuracy does not always require the highest latency. NeRF-FusionNet outperforms DETR-3D in accuracy while maintaining lower latency. This observation indicates that hybrid architectures can achieve efficient performance optimization. The observed trade-off pattern highlights the importance of evaluating accuracy and latency simultaneously. For time-critical systems, YOLOv8-3D remains a viable option despite its lower accuracy. For accuracy-critical tasks, NeRF-FusionNet offers a more suitable solution. DETR-3D represents a middle-ground approach with balanced accuracy and computational cost. Overall, Figure 3 supports informed decision-making for 3D object detection system design and deployment.

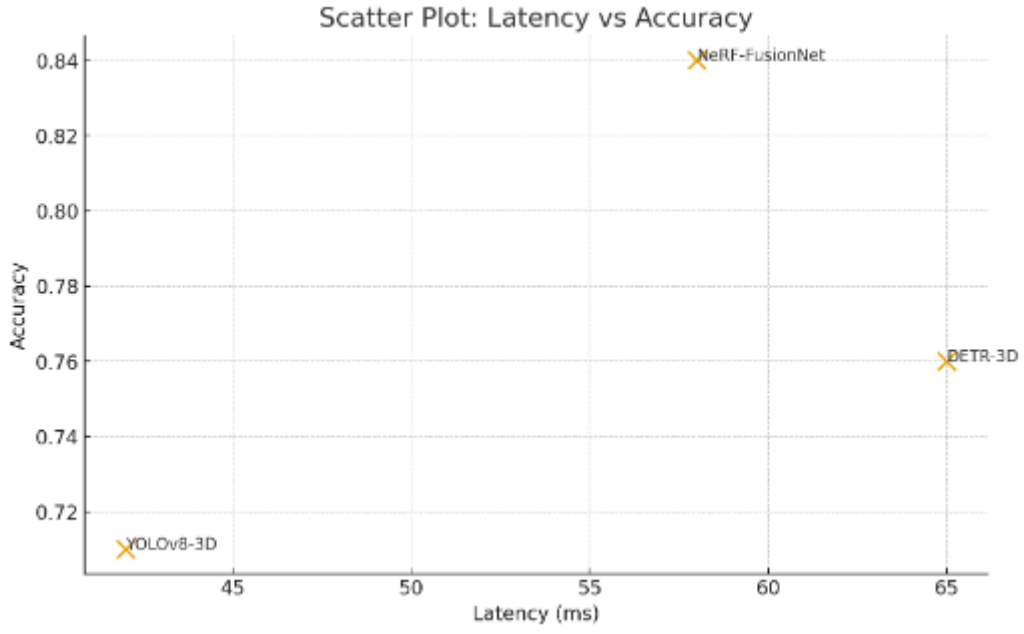

Figure 3. Latency-accuracy trade-off among 3D object detection models

Figure 4 illustrates the distribution of memory usage among the evaluated 3D object detection models using a box plot representation. The vertical axis shows memory consumption measured in megabytes. The box plot summarizes the central tendency and variability of memory usage across different models. The median memory usage is indicated by the horizontal line inside the box. This median lies in the upper-middle range of the overall distribution. The interquartile range reflects

differences in architectural complexity. Models with simpler structures tend to consume less memory. In contrast, more advanced models require higher memory resources.

The lower whisker of the box plot extends to approximately 510 MB, indicating the minimum observed memory usage. This value suggests that some models are optimized for lower resource environments. The upper whisker reaches nearly 840 MB, representing the maximum memory consumption. Such high usage is typically associated with models that employ richer feature representations. The wide range between the whiskers demonstrates substantial variability across models. This variability reflects differences in parameter size and internal processing mechanisms. Memory-intensive models often support more complex spatial reasoning. However, they may be less suitable for deployment on limited hardware.

From a practical perspective, Figure 4 highlights the importance of memory considerations in model selection. High memory consumption can limit deployment on edge devices or embedded systems. Conversely, models with lower memory requirements are more flexible for real-time applications. The box plot also helps identify potential outliers in memory usage. Understanding this distribution supports balanced decision-making in system design. Memory efficiency should be evaluated together with accuracy and latency metrics. Trade-offs between performance and resource usage are unavoidable in practice. Overall, Figure 4 provides a clear overview of memory usage characteristics for 3D object detection models.

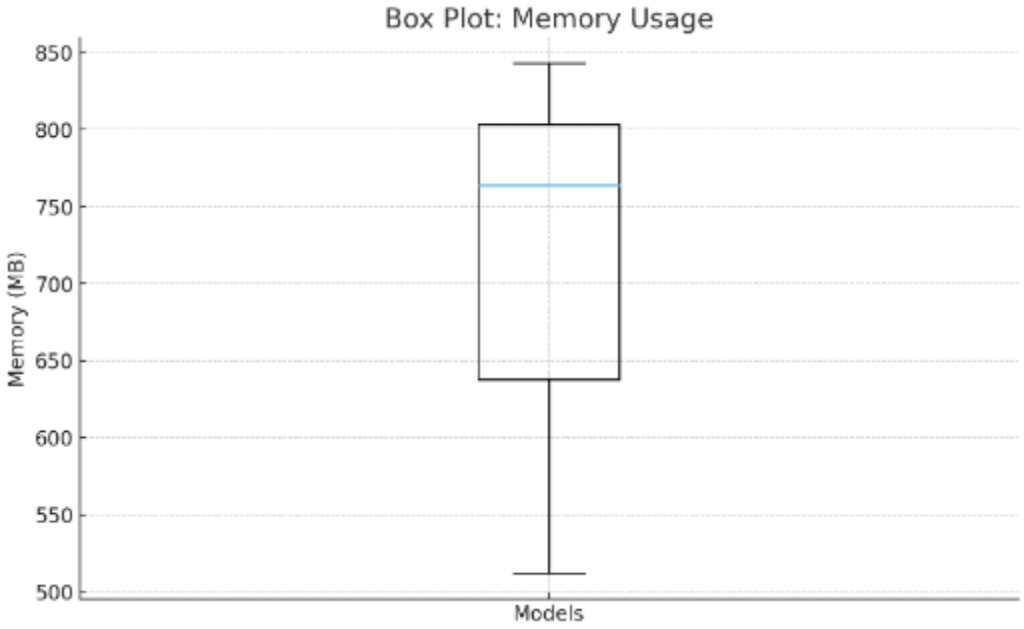

Figure 4. Distribution of memory usage across 3D object detection models

Figure 5 presents a heatmap that compares the performance of three 3D object detection models across multiple evaluation metrics. The horizontal axis represents the models, namely YOLOv8-3D, DETR-3D, and NeRF-FusionNet. The vertical axis lists the evaluated metrics, including accuracy, 3D Intersection over Union (IoU), normalized latency, and normalized memory usage. Color intensity in the heatmap reflects the relative magnitude of each metric value. Darker shades indicate higher values, while lighter shades represent lower values. This visual encoding enables rapid comparison across both models and metrics. The figure provides a compact summary of multidimensional performance characteristics. Overall, the heatmap facilitates intuitive interpretation of trade-offs among competing models.

In terms of accuracy, NeRF-FusionNet achieves the highest score, followed by DETR-3D and YOLOv8-3D. A similar trend is observed for the 3D IoU metric, where NeRF-FusionNet shows a substantial advantage. These results indicate stronger spatial localization and object representation capabilities. YOLOv8-3D records the lowest values for both accuracy and IoU, reflecting its emphasis on speed-oriented design. DETR-3D occupies a middle position, benefiting from transformer-based

Received: December 16, 2025. Revised: March 16, 2026.

global context modeling. The gradual increase across models suggests progressive improvements in representational capacity. The heatmap clearly highlights these performance gradients. This makes comparative assessment more efficient than isolated metric tables.

The normalized latency and memory metrics reveal additional trade-offs among the models. YOLOv8-3D shows the lowest normalized latency, confirming its efficiency in inference speed. However, this efficiency is accompanied by lower accuracy-related metrics. NeRF-FusionNet exhibits higher normalized memory usage, reflecting the cost of richer feature representations. DETR-3D demonstrates balanced values across latency and memory, positioning it as a compromise solution. The heatmap shows that no single model dominates all metrics simultaneously. Instead, each model exhibits strengths aligned with specific design priorities. Overall, Figure 5 provides a holistic view of performance trade-offs to support informed model selection.

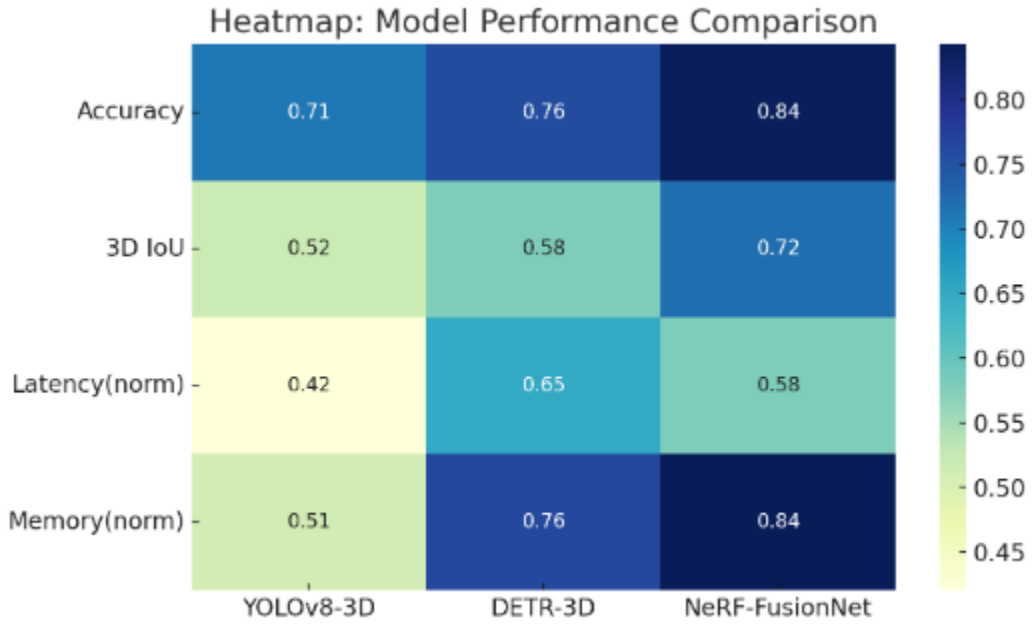

Figure 5. Heatmap-based comparison of model performance metrics

5. Conclusion

This study presented a comprehensive evaluation of three representative 3D object detection models—YOLOv8-3D, DETR-3D, and NeRF-FusionNet—by jointly analyzing detection accuracy, spatial alignment, inference latency, frame rate, and memory usage under realistic mobile deployment conditions. The results show that each model exhibits distinct strengths aligned with its architectural design. YOLOv8-3D consistently demonstrated the lowest latency and highest frame rate, making it well-suited for real-time applications on resource-constrained devices. DETR-3D provided a balanced performance profile, achieving improved spatial accuracy over YOLOv8-3D while maintaining moderate computational cost. NeRF-FusionNet achieved the highest detection accuracy and spatial alignment, indicating strong robustness in complex 3D environments, albeit with higher memory consumption.

The cross-platform evaluation on Snapdragon and Apple devices revealed consistent performance trends across hardware, suggesting stable generalization behavior of the evaluated models. While Apple platforms generally offered slightly better latency and frame rate, the relative ranking of models remained unchanged. These findings emphasize that platform optimization can improve efficiency but does not eliminate inherent architectural trade-offs. Importantly, the results highlight that high accuracy does not always require the highest latency, as demonstrated by the competitive efficiency of NeRF-FusionNet. This observation underscores the potential of hybrid architectures in balancing performance and computational demands.

Overall, this work demonstrates the necessity of multidimensional evaluation when selecting 3D object detection models for practical deployment. Focusing solely on accuracy or speed may lead to suboptimal choices for real-world applications. By integrating effectiveness and efficiency metrics,

Received: December 16, 2025. Revised: March 16, 2026.

this study provides actionable insights for system designers and practitioners. Future work may explore further optimization of hybrid models, energy consumption analysis, and evaluation on broader datasets and devices. Such extensions would strengthen the applicability of 3D object detection systems in diverse real-world scenarios.

References

[1] W. Wang, X. Li, X. Qiu, X. Zhang, V. Brusic, and J. Zhao, "A privacy-preserving framework for federated learning in smart healthcare systems," Inf. Process. Manag., vol. 60, no. 1, p. 103167, 2023, doi: https://doi.org/10.1016/j.ipm.2022.103167.

[2] M. Davis *et al.*, "OGITO, an Open Geospatial Interactive Tool to support collaborative spatial planning with a maptable," *Procedia Comput. Sci.*, vol. 227, no. 1, pp. 591–598, 2023, doi: https://doi.org/10.1016/j.geoforum.2023.103848.

[3] R. Das and M. M. Inuwa, "A review on fog computing: Issues, characteristics, challenges, and potential applications," *Telemat. Informatics Reports*, vol. 10, p. 100049, 2023, doi: https://doi.org/10.1016/j.teler.2023.100049.

[4] L. L. L. Zhang *et al.*, "Simulation of E-learning virtual interaction in Chinese language and literature multimedia teaching system based on video object tracking algorithm," *Expert Syst. Appl.*, vol. 10, no. 2, p. 102585, 2024, doi: https://doi.org/10.1016/j.jksuci.2020.08.007.

[5] A. Zadeh *et al.*, "The moderating effect of algorithm literacy on Over-The-Top platform adoption," *Comput. Human Behav.*, vol. 60, no. 3, p. 106403, 2023, doi: https://doi.org/10.1016/j.chb.2019.09.030.

[6] M. G. Hasan *et al.*, "Environmental monitoring with sensitive photonic fiber sensors for hazardous chemicals," *Int. Commun. Heat Mass Transf.*, vol. 162, p. 108631, 2025, doi: https://doi.org/10.1016/j.icheatmasstransfer.2025.108631.

[7] S. Sivanraj, D. N. L. Uduwage, and M. Tripathi, "Real-time safety detection on construction sites using a vision-language and NLP-based model," Adv. Eng. Informatics, vol. 69, p. 103889, 2026, doi: https://doi.org/10.1016/j.aei.2025.103889.

[8] Z. Xu, B. Wei, and J. Zhang, "Reproduction of spatial–temporal distribution of traffic loads on freeway bridges via fusion of camera video and ETC data," *Structures*, vol. 53, pp. 1476–1488, 2023, doi: https://doi.org/10.1016/j.istruc.2023.05.023.

[9] C. Esposito, V. Moscato, and G. Sperlì, "Detecting malicious reviews and users affecting social reviewing systems: A survey," *Comput. Secur.*, vol. 133, p. 103407, 2023, doi: https://doi.org/10.1016/j.cose.2023.103407.

[10] L. Diao, X. Tang, J. Wang, G. Xie, and J. Hu, "Hierarchical visual-semantic interaction for scene text recognition," *Inf. Fusion*, vol. 102, p. 102080, 2024, doi: https://doi.org/10.1016/j.inffus.2023.102080.

[11] H. Wang *et al.*, "Automated productivity analysis of cable crane transportation using deep learning-based multi-object tracking," *Autom. Constr.*, vol. 166, p. 105644, 2024, doi: https://doi.org/10.1016/j.autcon.2024.105644.

[12] S. Sarwar and A. G. L. Borthwick, "Estimate of uncertain cohesive suspended sediment deposition rate from uncertain floc size in Meghna estuary, Bangladesh," *Estuar. Coast. Shelf Sci.*, vol. 281, 2023, doi: 10.1016/j.ecss.2022.108183.

Received: December 16, 2025. Revised: March 16, 2026.