International Journal on Robotics, Automation and Sciences

# Deep Learning-based Obstacle Detection for Human Interaction Robots: A Review

Farhana Ahmed, Nor Hidayati Abdul Aziz*, Rosli Besar, Saad Salam and Md. Abdullah Man

***Abstract* – Obstacle detection is the foundation of autonomous robotics, enabling robots to perceive and understand the world around them to move safely. Deep learning has emerged as one of the driving forces in today's research, with various algorithms employed for learning and making effective decisions based on vast and complex datasets. In recent years, numerous deep learning methods have been developed and studied to detect obstacles. This paper provides an end-to-end overview of over 40 state-of-the-art deep learning models (from 50 papers) for obstacle detection in human-interacting robots, with a focus on deployment viability, real-time running, and energy efficiency. We also delve into the architecture of deep learning, highlight key challenges in real-world deployment, offer a comparative analysis of basic and advanced deep learning approaches, and examine the trade-offs between accuracy, speed, and power consumption, providing insights into practical considerations. This review categorizes obstacle detection techniques into two groups: Core CNN-based methods and Advanced Deep Learning Methods. Comparisons were made between these two groups, concentrating on computational requirements, deployment feasibility, and hardware configuration. Several key findings emerged. It was determined that models with high accuracy were computationally expensive and unsuitable for embedded deployment. While some models experience accuracy-speed trade-offs, others are limited by hardware constraints and power limitations. Finally, this review concludes with a structured discussion of real-world deployment considerations, prioritizing model efficiency, scalability, and potential future research directions in deep learning-based obstacle detection.**

***Keywords*— *Obstacle Detection, Deep Learning, Power Consumption, Computational Efficiency, CNN-Based Method, Deployment Feasibility, Lightweight Model.***

## I. INTRODUCTION

In the evolving environment of autonomous systems, human-interacting robots have emerged as key tools in diverse applications, from healthcare [1] to warehouses [2], education, home care, and the smart industry. To collaborate with humans efficiently and safely, these robots are outfitted with strong and accurate obstacle detection (OD), a capability required for autonomous navigation and real-time navigation in dynamic and largely unstructured environments. Obstacle detection was previously achieved through rule-based algorithms, handcrafted feature extraction [3], and sensor fusion [4] (i.e., LiDAR, ultrasonic, and infrared sensors). While these approaches provided baseline solutions, they lacked adaptability, generalizability to diverse environments, and real-time processing efficiency.

The shift to data-driven approaches was a fundamental inflection point in obstacle detection. Through the advent of deep learning and, in particular, Convolutional Neural Networks (CNNs) [5], obstacle detection shifted from hand-designed features

*Corresponding Author email: hidayati.aziz@mmu.edu.my, ORCID: 0000-0001-7995-4912
Farhana Ahmed, Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia (e-mail: farhanaahmad2000@gmail.com).
Nor Hidayati Abdul Aziz, Centre for Advanced Analytics, CoE for Artificial Intelligence, Multimedia University, Melaka, Malaysia. (e-mail: hidayati.aziz@mmu.edu.my).
Rosli Besar, Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia (e-mail: rosli@mmu.edu.my).
Saad Salam, Mechanical Engineering Department, Chittagong University of Engineering and Technology, Bangladesh (e-mail: saadsalam704@gmail.com)
Abdullah Man, TM R&D, Cyberjaya, Malaysia (e-mail: abdullah@tmrnd.com.my)

International Journal on Robotics, Automation and Sciences (2025) 7, 3:75-86
https://doi.org/10.33093/ijoras.2025.7.3.10
Manuscript received: 30 Jun 2025 | Revised: 28 Jul 2025 | Accepted: 11 Aug 2025 | Published: 30 Nov 2025

Published by MMU PRESS. URL: http://journals.mmupress.com/ijoras

engineering to automatic feature learning, enabling more expressive interpretation of visual and spatial data. CNNs delivered state-of-the-art performance in object recognition, semantic segmentation, and depth estimation, all of which are highly critical for robust obstacle detection in robotics. Unlike traditional methods, which involved the explicit programming of visual cues, CNNs can learn to detect obstacles directly from raw sensor data, significantly enhancing detection accuracy and robustness.

Following CNN's success, newer architectures, such as Recurrent Neural Networks (RNNs), Transformer models, and Reinforcement Learning (RL), have further advanced robotic perception abilities. These sophisticated methods support object detection, context-aware navigation, and adaptive learning in unpredictable environments. The integration of deep learning with real-time processing platforms has also enabled the deployment of complex models in practical robotic systems, overcoming previous issues related to latency and computational overhead.

In light of the increasing complexity and deployment challenges in robotic perception, this study aims to address several key research questions. First, it examines which deep learning architectures offer the best balance among inference speed, detection accuracy, and resource efficiency for indoor human-interaction robots. Second, it assesses how current state-of-the-art approaches, such as attention mechanisms, transformer models, and hybrid CNN-RNN architectures, compare to traditional CNNs in terms of real-time deployability. Finally, this paper seeks to identify the most significant research gaps in the current literature regarding embedded deployment, generalization across datasets, and effective model optimization.

This review hence presents a new challenge-based taxonomy of deep-learning-based obstacle detection methods to explore these questions systematically:

Core CNN-Based Methods, including real-time object detection models (YOLO, SSD), feature extraction networks, and custom CNN architectures. Advanced Deep learning Techniques, covering attention mechanisms, hybrid CNN-RNN models, and multi-task learning systems.

This paper covers algorithmic advances and emphasizes real-world deployment considerations such as speed–accuracy trade-offs, power budgets, hardware compatibility, and model size. Models are systematically benchmarked on accuracy, inference speed (FPS), and edge-device feasibility. The key contributions of this article are as follows:

A novel categorization of two major areas, Core CNN (deployment-focused) and Advanced DL (innovation-focused), enabling targeted evaluations for robotic applications. This review provides a systematic comparison of models based on accuracy, speed, hardware feasibility, and power efficiency, which are often overlooked in previous reviews. A significant contribution lies in its systematic comparative benchmarking of deep learning models on accuracy, speed (FPS), model size (MB), and embeddable device compatibility, addressing basic weaknesses like data scarcity, ineffective computations, and sparse transfer learning discovery, identifies lightweight architectures and optimization methods (e.g., quantization, pruning) for embedded systems.

By addressing these less-addressed dimensions, this paper aims to guide the design of effective, flexible, and deployable obstacle detection systems for human-interactive robots.

The following sections of this paper are organized: Section II, the background study; Section III, the literature review; Section IV, discussion and analysis findings; and Section V, conclusion.

## II. BACKGROUND & TECHNICAL FOUNDATION

### *A. Object Detection & Role in Human Interaction Robots*

Obstacle detection [6] is a crucial function in autonomous robotic systems, enabling robots to perceive their environment and avoid collisions during navigation. Unlike generic object detection, which involves recognizing objects belonging to known classes, obstacle detection specifically targets any object that could obstruct a robot's path, regardless of its semantic classification. These objects can include static obstacles (e.g., furniture), dynamic obstacles (e.g., pedestrians, oncoming vehicles), and environmental hazards (e.g., stairs, slopes).

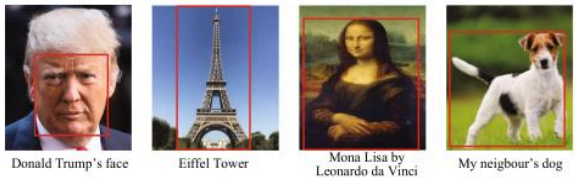


**FIGURE 1. Object Detection of a Particular Object** [6]

### *B. Core Components of Deep Learning Pipelines for Obstacle Detection*

A typical deep learning pipeline for obstacle detection involves several key stages, as shown in Figure 2:


**FIGURE 2. Obstacle Detection Pipeline**

*Step 01: Preprocessing of Input*

Raw sensor data (images, depth maps, LiDAR) are normalized, resized, and sometimes fused to yield a homogeneous input representation.

*Step 02: Feature Extraction*

Hierarchical features are learned by Convolutional Neural Networks (CNNs) from the input image. These features capture edges, textures, and higher-level patterns that are useful for obstacle detection.

*Step 03: Detection Head*

Generates proposal regions (in two-stage detectors like Faster R-CNN), or Suction bounding boxes and class probabilities tightly (in one-stage detectors such as YOLO or SSD).

*Step 04: Post-processing*

Non-Maximum Suppression (NMS) methods remove duplicate detections, and confidence thresholds remove low-probability predictions.

The pipeline enables robotic systems to process visual data in close to real-time, making it suitable for applications where real-time decision-making is necessary, such as in crowded human environments.

### *C. Problems in Object Detection*

Detecting small objects remains a challenge due to limited pixel representation and loss of context in deep CNNs. Although models like Faster R-CNN and YOLO are inefficient at detecting small objects, this issue can be addressed by increasing input resolutions, using data augmentation, and applying attention mechanisms to preserve fine details. However, recent advances have attempted to overcome this limitation by increasing input image resolution, employing data augmentation techniques, and leveraging attention mechanisms to preserve the detail of small objects better. Apart from the accuracy problems, efficiency and scalability are also ongoing issues. As the number of object classes grows, the computational cost of detection models grows as well, requiring more resources and reducing scalability. To address the annotation cost in big data sets, weakly supervised [7] learning techniques have been explored, which allow for more real-world model development. On top of that, optimizing detection models [8] to have scalability and efficiency in deployment has been a significant research area. Environmental robustness is another significant concern, as current best practices do not handle dynamic scenes with walking pedestrians or changing lighting conditions. These models, usually optimized for static scenes, lose accuracy when presented with motion blur or occlusion. Conversely, models optimized for dynamic scenes compromise on accuracy when applied to static scenes. These challenges highlight the prevailing trade-offs and compromises in achieving robust, real-time object detection in diverse indoor and outdoor settings.

### *D. Common Architectural Paradigms Used in Obstacle Detection*

A wide range of deep architectures powers modern obstacle detection models, each optimized with specific trade-offs in mind, such as accuracy, inference rate, memory consumption, and energy usage. With autonomous robots now being increasingly utilized in real-world human-interacting environments, it is necessary to understand these trade-offs in selecting the most suitable deployment model. We categorize the well-known architectures into four paradigms in Table 1 and provide a commentary on their relative suitability for real-time obstacle detection, accompanied by a flowchart.

**TABLE 1. Comparison of modern object detection model architectures, highlighting their key strengths and limitations.**

| Model Type | Description | Strength | Limitation |
|---|---|---|---|
| Two-Stage Detectors (Faster R-CNN, Mask R-CNN) [9] | Use region proposal networks followed by classification and regression heads. | Excellent accuracy | Slow inference |
| One-Stage Detectors (YOLO, SSD) [10] | Predict bounding boxes directly from feature maps. | Fast inference | Slightly lower accuracy |
| Lightweight CNNs (MobileNet, EfficientNet) [11] | Optimized for low-power devices. | Energy-efficient | May sacrifice accuracy |
| Attention-based Models / Transformers (DETR, Swin Transformer) [12] | Use global attention mechanisms for better contextual understanding. | Strong generalization | Memory-intensive |

These models aim at different deployment needs. For instance, YOLO-like models have been widely deployed on mobile robots due to their low latency, while Transformer-based models have future deployment prospects for interpreting complex scenes. Here is a flow chart added in Figure 3 to clarify the major components of detection pipelines and outline the leading architectural paradigms.

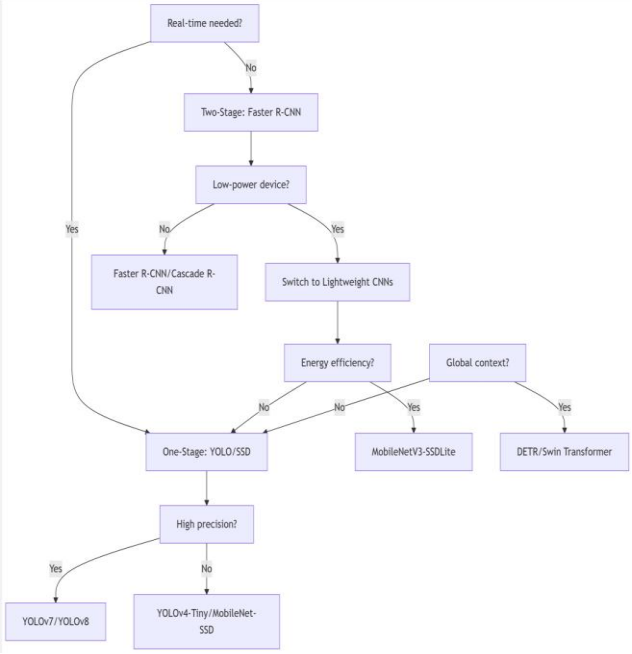


**FIGURE 3. Model selection flowchart for object detection, illustrating the trade-offs between two-stage, one-stage, lightweight, and transformer-based architectures.**

## III. CHALLENGE-DRIVEN LITERATURE REVIEW

Deep learning is currently the backbone of robotic perception, particularly in obstacle detection in human-interaction robots such as service, assistive, and collaborative indoor robots. These robots need proper perception but also real-time feedback, energy efficiency, and the ability to execute on low-power embedded platforms. This part provides a problem-driven taxonomy of deep learning techniques and a comprehensive survey of prevailing methods along the axes of deployment feasibility, computational compromises, and scope of use in human-robot interaction (HRI) setups.

### *A. A New Taxonomy: Challenge-Driven Categorization*

To provide practical insights, we categorize the reviewed approaches based on practical difficulties in real-world scenarios rather than merely algorithmic families. Table 2 lists the main deployment issues of Human Interactive robots and maps each to the respective deep learning paradigms and exemplar models.

**TABLE 2. Deep learning paradigm vs deployment challenges**

| Deployment Challenge | DL Paradigm | Representative Models |
|---|---|---|
| Real-Time Obstacle Detection | Core CNN (YOLO, SSD) | YOLOv3/v5, SSD, Faster R-CNN |
| Embedded Platform Compatibility | Lightweight CNN | YOLOv4-Tiny, MobileNet, RH-Net |
| Environmental Complexity | Transformer-Based Architectures | ST-YOLOX, RH-Net, ViT |
| Dynamic Scene Understanding | CNN + RNN/LSTM Hybrids | RFA-Net, MobileNet+LSTM |
| Long-Term Autonomy | Deep Reinforcement Learning | ResNet18 + YOLOv3 + PID |
| Generalization Across Environments | Self-Supervised & Multi-Task Learning | ST-YOLOX, ViT, KITTI & Marine Datasets |
| Efficiency Frontier | Neural Architecture Search (NAS) | PSO-NAS, GC-YOLO, Diff. NAS |
| Multimodal Robustness | Cross-Sensor Fusion | SparseFusion, Radar-Cam Polar Fusion, Robust-FusionNet |
| Hardware-Aware Co-Design | Transformer Optimization | HM-ViT, BiViT, V2X-ViTv2 |

This challenge-based classification allows us to compare models not merely according to precision, but also according to the feasibility of application, which is fundamental for real-world robot deployment.

### *B. Architectural Evolution Over Time (2021–2025)*

To further illustrate the evolution of deep learning architectures for obstacle detection, Figure 4 presents a timeline of significant architectural advancements from 2021 to 2025. This figure captures the evolution of models addressing specific deployment challenges, such as real-time performance, embedded support, and environmental complexity.

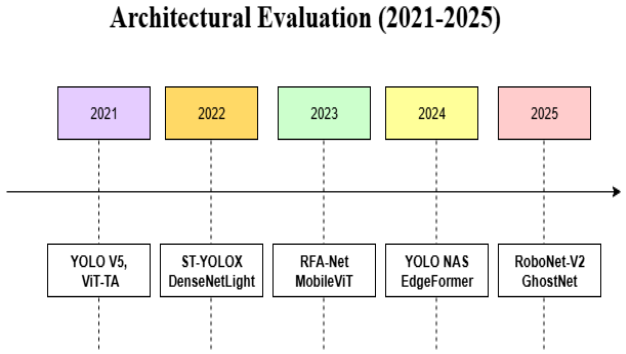


**FIGURE 4. Architectural evolution of deep learning model (2021–2025)**

Figure 4, the architectural scene has dramatically evolved in the past few years: **2021:** Early models like YOLOv5 and ViT-TA formed the foundation for high-performance obstacle detection, struck with computational efficiency. **2022:** The introduction of ST-YOLOX and DenseLightNet indicated significant breakthroughs in being able to handle environmental complexity and real-time performance, particularly for edge devices. **2023:** Technologies such as RFA-Net and MobileViT, targeted at enhancing embedded platform support and energy efficiency, positioned them well for low-power devices. **2024:** Methods such as YOLO-NAS and EdgeFormer demonstrated the ability of neural architecture search (NAS) and multimodal fusion in achieving accuracy and deployability. **2025:** New architectures like RoboNet-V2, GhostNet are pushing limits for dynamic scene understanding and long-term autonomy, integrating temporal information and reinforcement learning for better obstacle detection.

This evolutionary trend emphasizes the successive enhancement of deep learning models to respond to the diverse deployment challenges outlined in Table 2. Through dissecting these enhancements, we can have a better appreciation of the accuracy against speed against resource utilization trade-offs, and where future research potential lies.

### *C. Deep Learning Model by Deployment Challenge*

***Real-Time Obstacle Detection (Core CNN Models):*** Traditional Convolutional Neural Networks (CNNs) remain central to robotic vision due to their simplicity and performance. YOLOv3, YOLOv5, SSD, and Faster R-CNN have been demonstrated with excellent capabilities for obstacle detection with varied accuracy-speed trade-offs.

SSD-MobileNetV2, employed by Udink et al. [13] provides high FPS and reasonable accuracy in an organized indoor area, but comes at the cost of decreased performance in harsh lighting and clutter. Prakhar [14] contrasted YOLO and Faster R-CNN and favored YOLO for use on an embedded system. Abhinav et al. [15] also, reasserted that YOLOv5 was more effective for real-time support in assistive navigation. Despite satisfactory real-time performance, these models are less adaptive to dynamic lighting and occlusion, which limits their generalization.

***Insight***: While these models are good baselines for indoor perception, they require robust additions like

attention modules or multimodal fusion to handle human-rich environments.

***Embedded Platform Compatibility (Lightweight CNN Models):*** Low-cost Human-interaction robots must operate on energy-efficient embedded devices. This has led to the development of lightweight CNNs with decent accuracy and reduced computational complexity.

RH-Net, proposed by Zongyang et al. [16] is a CNN-Transformer hybrid used for real-time detection of railways. It operates at 43 FPS and 97% accuracy on Jetson Xavier NX. Jenefa [17] a DCNN method combining MobileNetV2 and LSTM for railway track obstacle detection, achieving high accuracy but limited by high computational demands and inefficiency on low-power hardware. Mouna et al. [18] utilized a modified YOLOv3 with MobileNet V1 for indoor object detection, offering compact and fast performance but struggling with object occlusion and limited generalizability. YOLOv4-Tiny and YOLOv5-Lite reduce model layers and complexity while preserving detection speed. YONG LV et al. [19] achieved real-time detection on sweeping robots using YOLOv4-Tiny. Ling et al. [20] proposed a lightweight three-stage detection framework for railway obstacle detection, optimizing YOLOv4-Tiny for low-power GPUs but suggesting further fine-tuning for ultra-low-power devices.

***Insight***: Lightweight CNNs are the most promising solution currently available for real-time, embedded obstacle detection, especially when Model pruning and quantization have been used for optimization.

***The Efficiency Frontier: Ultra-Lightweight Architectures:*** Lightweight model design has passed through three explicit phases:
(1) Hand-crafted efficient networks (2017-2020),
(2) Post-training compression (2020-2022), currently
(3) Natively efficient architectures (2023-present).

The long-standing issue of balancing efficiency in computation for accuracy has entered an era of revolution. Where aggressive pruning and quantization were a requirement up until now, next-generation lightweight models now leverage neural architecture search (NAS) and structural innovation to break this tradeoff*.*

Tao Gong, Yongjie Ma [21] proposes a PSO-optimized NAS architecture for fast object detection, using Ghostconv modules and multi-objective optimization (accuracy/FLOPs/parameters) to achieve 17.01% mAP on VisDrone2019 in 0.6 GPU days. Dynamic early stopping and elite weight sharing are proposed in the method to practically balance detection performance with resource constraints (13.1M params, 57.2G FLOPs). This paper [22] demonstrates that 8-bit quantized SSD MobileNetV2 achieves close-to-original accuracy (80.65% vs. 81.72%) but considerably improves edge-device efficiency (6ms latency, 334.6KB RAM on Raspberry Pi 5). The corresponding model excels in masked-face detection (F1=0.92) and performs best on Raspberry Pi 5/Jetson Orin Nano but fails on microcontrollers (>1s latency). GC-YOLO proposes [23] a lightweight YOLOv5 model for wheelchair blind-spot detection with 38% fewer parameters and 49.7% fewer GFLOPS, and higher mAP (84.19%→90.34%) and F1-score (0.62→0.84) via GhostConv and CoordAttention optimizations. The model runs in real-time at 24 FPS, providing a lighter substitute for computation-heavy LiDAR/3D approaches in semi-enclosed spaces. Amrita Rana and Kyung Ki Kim [24] presents a computation-efficient object detection NAS architecture that combines weight-sharing backbones with differentiable search, with 64.4% mAP (PASCAL VOC) at 4.8M parameters. Depthwise convolutions in the design and constrained search space reduce both the complexity of the model (1256M FLOPS) and search time (50% less than DARTS), and it becomes edge deployable.

***Insight***: NAS and architectural innovations (not just quantization) are driving the lightweight revolution.

***Environment Complexity (Transformer-Based Method):*** Indoor robots are often confronted with changing light, occlusion, and clutter problems that cannot be addressed by CNNs alone. Transformer-based architectures address this through global spatial feature learning and long-range dependencies.

DenseLightNet utilizes group convolutions for low overhead but is still optimized for edge deployment [25]. ST-YOLOX, which employs Swin Transformers for object contextual detection. It achieves 89–94% accuracy at the expense of high computational power [26]. Vision Transformers (ViT-TA) [27] used for traffic condition classification, which also achieves higher accuracy but is plagued with quadratic memory growth on embedded systems.

***Insights:*** These models are highly accurate but are not suitable for low-cost HRI robots without being optimized by hardware-aware distillation or edge-adaptive attention mechanisms.

***Dynamic Scene Understanding (Hybrid CNN +RNN/LSTM):*** To simplify temporal changes and object trajectories, some systems combine CNNs with RNNs or LSTMs:

RFA-Net [28] combines CNNs with memory-aware architectures and global distillation of expertise, with equilibrium performance (92.7% mAP, 40.4 FPS) but larger DRAM usage. Abdurrahman et al. [29] designed a robot named Lintang, which is significant for health issues. This paper utilized an LSTM algorithm to classify stunting based on age and height, achieving a great accuracy of 96.61% after training for 50 epochs.

***Insight***: These models are suitable for dynamic environments (e.g., navigation through crowds), but sequential processing leads to latency and power inefficiencies.

***Long-Term Autonomy (Deep Reinforcement Learning****):* Deep Reinforcement Learning (DRL) enables robots to learn adaptive obstacle avoidance strategies:

ResNet18 + YOLOv3 + PID controller, used by fang and Cai [30], enables low-cost autonomous navigation using Jetson Nano. As trivial as it appears to be, PID parameter tuning is still a bottleneck.

***Insight***: DRL methods are feasible for closed-loop control but require large-scale training and tedious tuning, hindering swift deployment.

***Generalization Across Environments (Self-Supervised and Multi-Task Learning):*** Both self-supervised and multi-task learning methods attempt to address critical challenges by training on diverse datasets captured in Figure 5, but the outcome is still not uniform across scenarios.

KITTI and Private datasets [31] achieve strong performance (up to 96%) in balanced scenarios with equitable lighting variation, but lack good 3D scene understanding since they are monocular depth estimation-based. Marine datasets [32] with 91% accuracy yield high-resolution images but struggle at pixel-level recognition of shapes due to water-induced noise, like reflection and turbidity. The MaSTr1325/MODD2/SMD/MID group's [33] performance is poor (60.08%) due to invalid validation processes and unbalanced lighting variability. VIT's Driving Dataset [34] shows excellent lane detection (91.36%) but drops to 75.28% accuracy in obstacle detection when tested in unstructured environments like fog and occlusion, demonstrating structural bias.

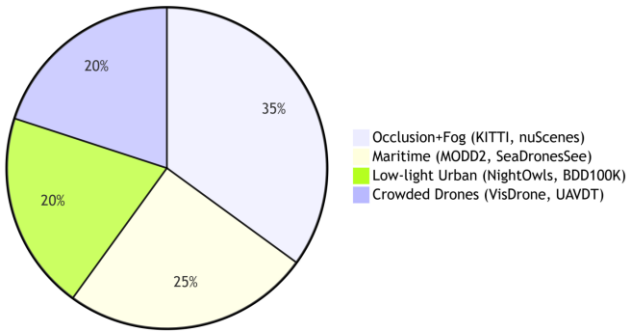


**FIGURE 5. Benchmark Dataset Details**

***Insight:*** Future models must adopt multi-modal learning and domain adaptation techniques to generalize well to novel situations***.***

***Multimodal Fusion for Real-World Robustness:*** The intrinsic constraints of single-sensor perception, whether it RGB's vulnerability to lighting changes, LiDAR's sparse depth data limitation, or event cameras' lack of texture, have had a long way to restrict obstacle detection in dynamic scenes.

SparseFusion introduces [35] a novel sparse-representation paradigm for efficient multi-sensor 3D object detection with parallel LiDAR/camera detectors and cross-modality transfer. It achieves state-of-the-art 73.8% NDS and 72.0% mAP on nuScenes at 5.6 FPS (1.8× faster than the prior work). The method demonstrates particular robustness to sensor outages, outperforming dense methods by +24.2% mAP in low LiDAR field conditions. Tianbi et al. [36] presents a decision-level fusion algorithm that unifies millimeter-wave radar and camera data in polar representation to enhance object detection for autonomous driving. Fusing YOLOv5-based visual detection with radar data, the system achieves 99.9% accuracy and 93.2% recall while reducing false alarms to 0.12%. Real-time efficiency (28 fps on Jetson TX2) makes it viable for edge deployment on smart vehicles.Robust-FusionNet [37] demonstrate a multimodal 3D object detection network for LiDAR and camera fusion to compensate for weather distortions. Its key contributions, pointwise alignment (K-means++), implicit feature pyramids (i-FPN), and hybrid attention (HAM), enhance accuracy by 2.59% on ONCE and 1.04% on KITTI datasets without trading real-time performance. The system is highly robust in adverse weather, with ablation studies confirming each contribution to the 12.59% aggregate accuracy improvement over baselines.

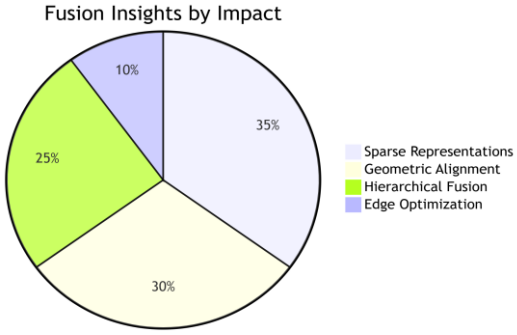


**FIGURE 6. Relative contribution of the four key technical approaches**

***Insight:*** Multimodal fusion (LiDAR/camera/radar) with cross-sensor alignment and sparse representation significantly increases accuracy and robustness in 3D object detection with real-time efficiency, especially in harsh conditions***.***

***Hardware-Aware Architecture Co-Design***: HM-ViT [38] presents the first homogeneous framework for hetero-modal V2V collaboration, enabling various sensor-equipped cars to collaboratively perceive 3D scenes via a novel heterogeneous 3D graph transformer. The framework couples local and global attention mechanisms ($H^3GAT$) with low-bandwidth feature sharing, achieving 8.8% improved AP@0.7 over homogeneous methods. OPV2V experiments show camera agents achieve 23x performance gains upon fusion with LiDAR data, upholding its real-world scalability. Another approach BiViT [39] introduces three breakthroughs to binarize Vision Transformers effectively:

Softmax-aware Binarization (SAB), reducing attention quantization error by 99.9% compared to BiBERT. Cross-layer Binarization (CLB), preserving pre-trained knowledge with a better accuracy of 11.6%, and Parameterized Weight Scales (PWS), improving module performance.

The method achieves 75.6% Top-1 accuracy on ImageNet with Swin-S (32x model compression) and 4.39x speedup of inference while maintaining 40.8 mAP on COCO detection. These advancements overcome the primary challenges in softmax binarization and data preservation and make ViTs suitable for edge deployment. V2X-ViTv2 [40] illustrate an advanced transformer-based cooperative perception framework for autonomous vehicles with

multi-scale pooling attention (MSPA) and robust fusion architectures. It is 5.8% more accurate in terms of AP over V2X-ViTv1 on V2XSet and beats the state of the art by 14.8% on DAIR-V2X while having stable performance against communication delay (10% accuracy decline at 400ms) and sensor noise. The innovation of the framework's attention mechanisms and V2X data augmentation techniques significantly enhances real-world multi-agent 3D object detection.

***Insight:*** Transformer advances in V2X (HM-ViT/V2X-ViTv2) and binarization (BiViT) bring real-world-capable multi-agent perception with robust fusion and edge optimization.

### *D. Cross-Dimensional Comparative Evaluation*

While the earlier sections describe deep learning models by deployment challenge, this section provides a technical comparison of selected models on quantifiable deployment criteria. Table 3 presents a model-by-model comparison on four crucial parameters: accuracy, inference rate (FPS), model size, and embedded design feasibility, to assess the suitability of these models for the requirements of real human-interaction robots. The table is meant to complement the algorithm-specific results by providing real, performance-based context.

**TABLE 3. Model performance vs deployment feasibility**

| Model | Accuracy | FPS | Model Size | Embedded Feasibility |
|---|---|---|---|---|
| YOLOv5 [19] | 95.2% | 15.9 | 1.4 MB | ☑ Excellent |
| RH-Net [16] | 97.0% | 43 | 14.1 GFLOPs | ☑ Yes |
| YOLOv4-Tiny [20] | 80% | 28 – 436 | Reduced Layers | ☑ Yes |
| ST-YOLOX [26] | 94% | 40 | High | ✗ No |
| ViT-TA [27] | 94% | - | Very High | ✗ No |
| RFA-Net [28] | 92.7% | 40.4 | Medium | ⚠ Needs tuning |

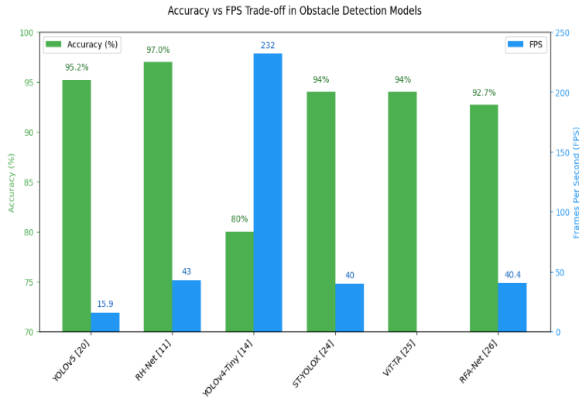


**FIGURE 7. Comparison of accuracy and frame-per-second (FPS) performance among different obstacle detection models**

Only a handful of models on the Figure.7 (YOLOv5-Lite, RH-Net) strike the right trade-off between real-time performance, accuracy, and resource usage, making them strong candidates for real-world human-interactive robot roll-out.

## IV. DISCUSSION AND ANALYSIS OF FINDINGS

### *A. Performance vs. Deployability: The Core Challenge*

Although deep learning revolutionized obstacle detection in human-robot interaction, not every high-performing model is deployable in practice. Being highly accurate is not enough; practical deployment for real-world use cases calls for lightweight models, low power usage, and edge device compatibility. In this review, models are sorted based on a new deployment-focused framework that aims for:

1. ***Lightweight:*** Architectures with low storage and compute footprints that make deployment feasible on resource-constrained edge devices without GPU capabilities.
2. ***Embedded-feasible:*** Tested on edge devices (Jetson, Raspberry Pi [41] ) with efficient architectures like MobileNet.
3. ***Conditional (Moderate):*** Require specific optimizations for low-power usage, but are feasible with adjustments.
4. ***Not Lightweight/Feasible:*** Large architectures needing a GPU or excessive memory for real-world testing.

### *B. Benchmark Comparison and Performance Highlight*

Through literature review-informed model-level benchmarking, this section offers a more compound synthesis by grouping algorithms into their respective model families, explicitly demarcated as Core CNN or Advanced Deep Learning (DL), and comparing across deployment-relevant dimensions like accuracy, inference speed, model size, and hardware feasibility. Rather than architecture-based individual approaches, Table 4 reflects more general trends of deployment and group-level performance characteristics using our proposed taxonomy: Lightweight, Embedded-Feasible, Conditional, and Not Feasible. The seven model families were not arbitrarily selected; instead, it is based on a structured review of the most prominent, most cited, and practiced proven classes in recent work in human-interaction indoor obstacle detection. They were chosen because they:

Capture prevailing deep learning paradigms (e.g., YOLO, MobileNet, Transformer) constantly being invoked in benchmark research. Exhibit architectural diversity across CNNs, hybrid models, attention-based networks, and reinforcement learning techniques. Provide empirical results that apply to real-world deployment, such as FPS, model compactness, energy efficiency, and embedded device compatibility.

The seven-model class taxonomy, therefore, forms an ideal, deployment-oriented foundation for comparative study and observation presented in this section. In Table 4, **[C]** denotes that the paper was collected from Core CNN models, and **[A]** depicts the paper used for the Advanced CNN Techniques.

**TABLE 4. Summary of object detection methods, their performance metrics, and deployment characteristics**

| Method | Accuracy | FPS | Lightweight | Embedded Device | Challenge |
|---|---|---|---|---|---|
| YOLO Based CNN **[C]** | 66.77 - 95.02% | 15 - 436 | **YES** (Tiny & YOLO v5) | **YES** (Jetson Pi) | A trade-off between speed and precision |
| SSD - Mobile Net V2, V3& CNN with RH-Net **[C]** | 79.80 - 97.02% | 30 - 135 | **YES** (<10-20 MB Model size) | **YES** (Jetson Xavier NX, GTX 1080Ti) | Well-suited for real-time on-edge devices. |
| Faster R-CNN & RCNN **[C]** | 66.00 - 69.00% | 5-17 | **NO** (FPS<10) | **Moderate** | low speed, GPU required. |
| Transformer Based **[A]** | 89.00 - 99.00% | 40-67 | **NO** (Large model size with 9.5 million parameters) | **NO** (Power Hungry) | Very accurate but compute-heavy, not for embedded. |
| Hybrid CNN + LSTM/ RNN **[A]** | 92.70 - 94.12% | 40.4 | **Moderate** (LSTM RAM/FLOPs overhead) | **YES** (Needs Xavier-class hardware) | Good accuracy, some need tuning. |
| Self / Multi-Task Learning **[A]** | 60.08 - 96.0% | 10+ | **YES** (Auto-encoders reduce parameters) | **Moderate** (Some models need GPUs) | Sensitive to environments. |
| DRL Model **[A]** | 96.01 % | - | **YES** | **YES** (Runs on Nano but platform-specific) | Requires PID tuning and is limited to certain platforms. |
| Ultra-lightweight **[A]** | 17.01-90.34% | 24-167 | **YES** | **YES** (runs on edge devices) | High FLOPs |
| Multi-model Fusion **[C]** | 2.59(mAp) – 99.9% | - | **Moderate** | **Moderate** | Sensor outages robustness (+24.2% mAP) |

In Figure 8, the pie chart summarizes the feasibility of deep model families based on their suitability for deployment. Every slice is a model family (such as YOLO-based CNNs or Transformers). Most are Lightweight or Embedded-Feasible, reflecting advancement in meeting real-time performance and efficiency needs for human-robotic interaction.

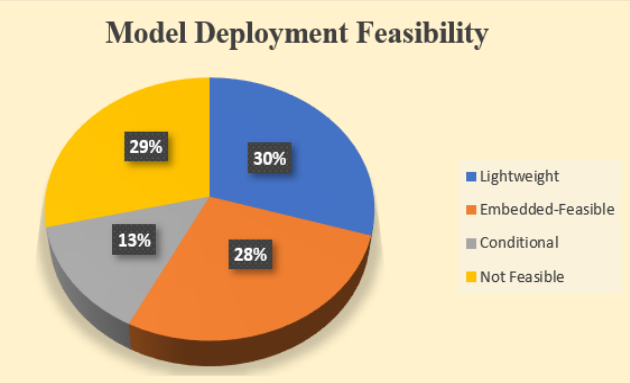


**FIGURE 8. Deployment Feasibility of Deep Learning Model Families.**

### *C. Scientific Insights: What the Performance Data Shows*

Base CNNs are increasingly faster and more efficient, enabling robot navigation within safely changing human environments. Hybrid CNN-LSTM architectures (e.g., RFA-Net) offer stronger temporal reasoning but at higher memory bandwidth and tuning complexity costs. Transformer-type solutions, though bound to generalize, incur high latency and computation burden, unacceptable for real-time service robots unless they undergo heavy optimization. To push the frontier, it is necessary to incorporate modular attention mechanisms, sensor fusion, and model pruning specific to the edge. In addition, standardizing benchmarks with unified datasets (e.g., KITTI, Marine, SMD) will be key for robust cross-environment verification.

This comparative benchmarking not only yields performance scores but also functional intelligence regarding where and how models can be deployed. It bridges the gap between raw data and deployable knowledge, a major novelty that this paper offers.

### *D. Research Opportunity to Fill the Gaps*

This review establishes many key research gaps that still remain unsolved in applying deep learning for real-world indoor obstacle detection, particularly in human-interactive robotics. These gaps are compiled by a concept map that is referenced in Figure 9 with a structured arrangement of key limitations under themes like dataset shortcomings, deployment inefficiencies, architectural flaws, and testing

constraints.

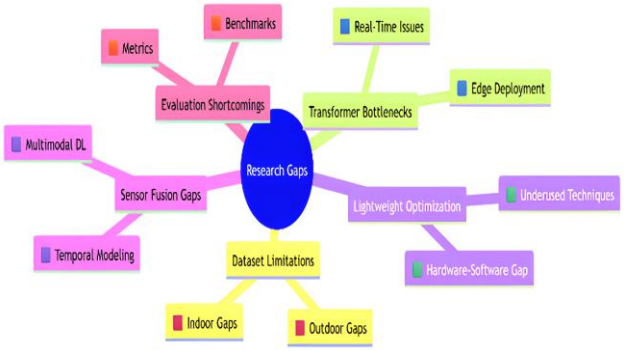


**FIGURE 9. Concept Map of Key Research Gaps in Deep Learning-Based Obstacle Detection.**

Figure 9 categorizes available research limitations into five classes, i.e., model optimization, benchmarks (transformer bottlenecks), dataset diversity, multimodal fusion, and real-time possibility. Every node represents an individual problem with direct applicability to real-world deployment. Based on our survey, the following gaps need to be addressed with all due urgency:

Most studies are based on outdoor/static information. There is no rich, real-time indoor dataset (e.g., crowded homes, varying lighting). Furthermore, techniques like pruning, quantization, and NAS are not fully utilized for transformer and hybrid models. A major limitation in current models is that they are not suitable for dynamic or occluded settings, and this highlights that real-time adaptive learning is needed. In addition, Fusing LiDAR, IR, and RGB in compact networks will enhance robustness. research into lightweight transformer architectures optimized for embedded inference is still in its early stages, and alternative solutions in this space remain scarce.

### *E. Alignment of Research Gaps with Initial Questions*

The highlighted research gaps have a direct relation and bear concrete answers to the initial research questions of this study.

Architectural trade-offs can be found in Table 4, which provides insight into the intricate balance in between speed, power consumption, and accuracy. This paper also answers Research Question 1 directly, showing that even though transformer-based models exhibit better accuracy, lightweight convolutional neural networks such as YOLOv5-Lite remain ahead in practical deployment, particularly within resource-constrained environments. This performance gap indicates the need for hardware-aware NAS, as discussed in Table 5, to be more appropriate in catering to model design in alignment with real-world hardware limitations. Besides, Research Question 2 (RQ2) concerns real-time deployability, and this is under threat from the inadequate pruning techniques typically applied to transformer models. Table 5 shows that even newer models like ST-YOLOX cannot handle real-time requirements, which reflects the need for attention mechanisms with an edge boost that might increase inference speed without compromising precision. Concluding, regarding Research Question 3 (RQ3), the unfavorable generalization performance shown by the current models on multiple domains in Table 3 and Figure 9 indicates the necessity of standardizing benchmark datasets and utilizing multimodal fusion techniques to boost the strength and adaptability of the models under novel and evolving indoor environments.

This consonance highlights the way our gap analysis (Figure 9, Table 5) methodically responds to the paper's starting questions and thus completes the loop between findings and motivation.

### *F. Gap Analysis and Opportunities*

To systematically tackle the constraints identified under Sections IV (C–E), we propose a comparative examination of existing problems and proposed solutions (Figure 10) and a comprehensive taxonomy (Table 5).

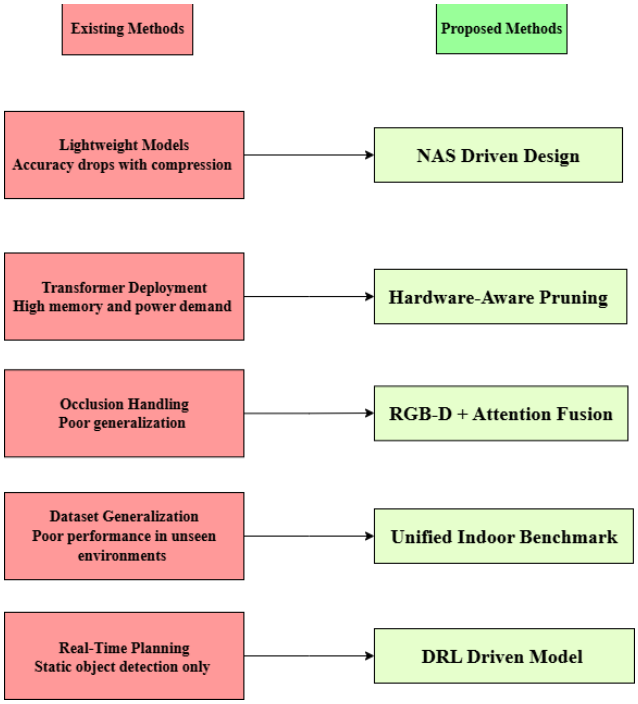


**FIGURE 10. Problem (Left side) – Solution (Right side) Relationship**

These gaps are also supported by quantitative findings in Table 5, where every challenge area is aligned with its respective limitation and potential future area of research. While deep learning has greatly advanced in terms of obstacle detection, there are certain critical limitations faced during porting such models to practical indoor robotic systems. Table V is a formal listing of such open issues and associated research opportunities. For example, lightweight models like YOLO that have an accuracy of 52 - 82% [42], where the YOLO NAS model secures an accuracy of 98.4% [43]. Their accuracy is compromised because of extreme compression; a research opportunity lies in examining the application of neural architecture search (NAS) towards creating small but accurate models. Similarly, transformer models, while powerful, are still too resource-hungry [44] for embedded platforms, necessitating hardware-aware pruning techniques [45].

Poor handling of occlusion [46] and dynamic scenes is another challenge to be addressed by the union of attention with multimodal sensor fusion [47] (e.g., combining RGB and depth information). Additionally, cross-domain generalization continues to be a concern owing to the absence of adequate benchmarking protocols [48] for indoor spaces, and hence the need for standard datasets and evaluation protocols. Finally, current approaches focus on static object detection [49] with little consideration for real-time motion planning, a task that can be approached by integrating detection and DRL-based avoidance techniques [50] in a single system.

**TABLE 5. Comparative Gap Analysis: Existing vs. Proposed Solutions**

| Challenge Area | Existing Limitations | Proposed /Research Opportunity |
|---|---|---|
| Lightweight Models | Accuracy drops (YOLO-Based Models) with compression [42] | Explore neural architecture search (NAS) for compact design [43] |
| Transformer Deployment | High memory and power demand [44] | Design hardware-aware transformer pruning [45] |
| Occlusion Handling | Poor generalization [46] | Combine attention with sensor fusion (RGB + Depth) [47] |
| Dataset Generalization | Poor performance in unseen environments [48] | Develop a unified benchmarking protocol across domains |
| Real-Time Planning | Static object detection only [49] | Integrate detection with DRL-based avoidance in a single model [50] |

Figure 9 and Table 5 together underscore the utmost significance of multi-disciplinary, co-operative research on deep learning-based model deployment-readiness in obstacle detection. Bridging these gaps will form the foundation of stronger, more scalable, and more flexible robotic systems for indoor service environments.

## V. CONCLUSION

This paper offered a comprehensive and deployment-focused review of deep learning-based obstacle detection approaches in human-interaction robots with a novel taxonomy that categorizes models into Core CNN-based and Advanced Deep Learning approaches. The comparison study in terms of accuracy, FPS, model size, and embedded deployment readiness offers practical guidance to developers seeking to deploy real-time and efficient systems on edge devices.

Compared to past reviews, which were either limited to accuracy or model categories only, this paper highlights real-world constraints like power efficiency, memory, and adaptability to dynamic environments. By hierarchically categorizing models based on their readiness for deployment from light-weight CNNs like SSD-MobileNetV2 to compute-intensive Transformer-based networks, the paper bridges a critical gap between theoretical development and practical robotic applications.

Despite significant advances, several challenges remain. One of the main research gaps is the lack of benchmarking datasets that accurately reflect the challenges of real-world indoor environments, such as varying lighting, occlusions, and human unpredictability. Another area of future research is the limited use of lightweight Transformers and attention-based models in embedded devices. Most current models, while accurate, continue to be constrained by resource demands that prohibit real-time deployment.

***Future Work:***

To transcend these limitations, future research must focus on:

Developing benchmarking datasets for indoor service and assistive robots in dynamic real-world environments. Designing and testing lightweight Transformer models specifically tailored for embedded devices using quantization, pruning, and neural architecture search. Exploring multi-modal sensor fusion methods (e.g., LiDAR, infrared, and vision) with deep learning for improved generalization and robustness. Investigating power-saving training paradigms like self-supervised or continual learning, which can facilitate adaptation through limited retraining of robots.

By matching model research to deployment realities and by targeting human-centric robotic applications, research in the future can render obstacle detection models not only accurate but also power-efficient, versatile, and deployable in indoor robotic systems in the real world.

## ACKNOWLEDGMENT

We thank the anonymous reviewers for the careful review of our manuscript. The authors also gratefully acknowledge the financial and technical assistance provided by Telekom Malaysia Research & Development (TMR&D) and Multimedia University (MMU), which enabled the completion of this study. This work is supported in part by the MMU-TMR&D Co-Creation Research Project, under Grant No-MMUE/240094.

## FUNDING STATEMENT

The funding for this work completion has been provided by the MMU-TMR&D Co-Creation Research Project, which is under Grant No. MMUE/240094.

## AUTHOR CONTRIBUTIONS

Farhana Ahmed: Literature search, conceptualization, drafting the methodology, concluding the research, and writing the manuscript;

Nor Hidayati Abdul Aziz: Critically reviewed, supervised, and provided feedback on the manuscript;

Rosli Besar: Critically reviewed, supervised, and provided feedback on the manuscript;

Saad Salam: Editing and refining the manuscript;

Md. Abdullah Man: Supporting the financial needs.

## CONFLICT OF INTEREST

No conflicts of interest were disclosed.

## ETHICS STATEMENTS

This research did not involve human participants, animal subjects or sensitive personal data and therefore did not require ethical approval.

## REFERENCES


[1] E.M.G.N.V. Cruz, S. Oliveira and A. Correia, "Robotics Applications in the Hospital Domain: A Literature Review," *Applied System Innovation 2024, Vol. 7, Page 125*, vol. 7, no. 6, p. 125, 2024. DOI: https://doi.org/10.3390/asi7060125

[2] H.A. Berkers, S. Rispens and P.M. Le Blanc, "The role of robotization in work design: a comparative case study among logistic warehouses," *International Journal of Human Resource Management*, vol. 34, no. 9, pp. 1852–1875, 2023. DOI: https://doi.org/10.1080/09585192.2022.2043925

[3] D. Ristić-Durrant, M. Franke and K. Michels, "A Review of Vision-Based On-Board Obstacle Detection and Distance Estimation in Railways," *Sensors 2021, Vol. 21, Page 3452*, vol. 21, no. 10, p. 3452, 2021. DOI: https://doi.org/10.3390/s21103452

[4] N.S. Ahmad, N.L. Boon and P. Goh, "Multi-sensor obstacle detection system via model-based state-feedback control in smart cane design for the visually challenged," *IEEE Access*, vol. 6, pp. 64182–64192, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2878423

[5] A.B. Atitallah, Y. Said, M.A.B. Atitallah, M. Albekairi, K. Kaaniche and S. Boubaker, "An effective obstacle detection system using deep learning advantages to aid blind and visually impaired navigation," *Ain Shams Engineering Journal*, vol. 15, no. 2, 2024. DOI: https://doi.org/10.1016/j.asej.2023.102387

[6] L. Liu *et al.*, "Deep Learning for Generic Object Detection: A Survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020. DOI: https://doi.org/10.1007/s11263-019-01247-4

[7] F. Shao *et al.*, "Deep Learning for Weakly-Supervised Object Detection and Localization: A Survey," *Neurocomputing*, vol. 496, pp. 192–207, 2022. DOI: https://doi.org/10.1016/J.NEUCOM.2022.01.095

[8] G. Li *et al.*, "Implicit Feature Contrastive Learning for Few-Shot Object Detection," *Computers, Materials and Continua*, vol. 84, no. 1, pp. 1615–1632, 2025. DOI: https://doi.org/10.32604/CMC.2025.063109

[9] L. Cao and X. Zhu, "An autonomous service mobile robot for indoor environments," *Proceedings of 2020 Asia-Pacific Conference on Image Processing, Electronics and Computers(IPEC) 2020*, pp. 8–15, 2020. DOI: https://doi.org/10.1109/IPEC49694.2020.9115180

[10] A. Younis, L. Shixin, J.N. Shelembi and Z. Hai, "Real-time object detection using pre-trained deep learning models mobilenet- SSD," *ACM International Conference Proceeding Series*, pp. 44–48, 2020. DOI: https://doi.org/10.1145/3379247.3379264

[11] M. Afif, R. Ayachi, Y. Said and M. Atri, "Deep embedded lightweight CNN network for indoor objects detection on FPGA," *Journal of Parallel and Distributed Computing*, vol. 201, p. 105085, 2025. DOI: https://doi.org/10.1016/J.JPDC.2025.105085

[12] C. Lin, Y. Cheng, X. Wang, J. Yuan and G. Wang, "Transformer-Based Dual-Channel Self-Attention for UUV Autonomous Collision Avoidance," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 3, pp. 2319–2331, 2023. DOI: https://doi.org/10.1109/TIV.2023.3245615

[13] U. Aulia, I. Hasanuddin, M. Dirhamsyah, and N. Nasaruddin, "A new CNN-BASED object detection system for autonomous mobile robots based on real-world vehicle datasets," *Heliyon*, vol. 10, no. 15, 2024. DOI: https://doi.org/10.1016/j.heliyon.2024.e35247

[14] P. Agrawal *et al.*, "YOLO Algorithm Implementation for Real Time Object Detection and Tracking," *2022 IEEE Students Conference on Engineering and Systems*, 2022. DOI: https://doi.org/10.1109/SCES55490.2022.9887678

[15] J.H. Kim, N. Kim, Y.W. Park and C.S. Won, "Object Detection and Classification Based on YOLO-V5 with Improved Maritime Dataset," *Journal of Marine Science and Engineering 2022, Vol. 10, Page 377*, vol. 10, no. 3, p. 377, 2022. DOI: https://doi.org/10.3390/JMSE10030377

[16] Z. Zhao, J. Kang, Z. Sun, T. Ye and B. Wu, "A real-time and high-accuracy railway obstacle detection method using lightweight CNN and improved transformer," *Measurement (Lond)*, vol. 238, 2024. DOI:https://doi.org/10.1016/j.measurement.2024.115380

[17] A. Jenefa *et al.*, "Real-Time Rail Safety: A Deep Convolutional Neural Network Approach for Obstacle Detection on Tracks," *ICSPC 2023 - 4th International Conference on Signal Processing and Communication*, pp. 101–105, 2023. DOI:https://doi.org/10.1109/ICSPC57692.2023.10125284

[18] M. Afif, Y. Said, R. Ayachi, and M. Hleili, "An End-to-End Object Detection System in Indoor Environments Using Lightweight Neural Network," *Traitement du Signal*, vol. 41, no. 5, pp. 2711–2719, 2024. DOI: https://doi.org/10.18280/ts.410544

[19] Y. Lv, Y. Fang, W. Chi, G. Chen and L. Sun, "Object Detection for Sweeping Robots in Home Scenes (ODSR-IHS): A Novel Benchmark Dataset," *IEEE Access*, vol. 9, pp. 17820–17828, 2021. DOI: https://doi.org/10.1109/ACCESS.2021.3053546

[20] L. Guan, L. Jia, Z. Xie and C. Yin, "A Lightweight Framework for Obstacle Detection in the Railway Image Based on Fast Region Proposal and Improved YOLO-Tiny Network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, 2022. DOI: https://doi.org/10.1109/TIM.2022.3150584

[21] T. Gong and Y. Ma, "PSO-based lightweight neural architecture search for object detection," *Swarm and Evolutionary Computation*, vol. 90, 2024. DOI: https://doi.org/10.1016/j.swevo.2024.101684

[22] H. Lokhande and S. R. Ganorkar, "Object detection in video surveillance using MobileNetV2 on resource-constrained low-power edge devices," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 1, pp. 357–365, 2025. DOI: https://doi.org/10.11591/eei.v14i1.8131

[23] J. Du, S. Zhao, C. Shang, and Y. Chen, "Applying Image Analysis to Build a Lightweight System for Blind Obstacles Detecting of Intelligent Wheelchairs," *Electronics (Switzerland)*, vol. 12, no. 21, 2023. DOI: https://doi.org/10.3390/electronics12214472

[24] A. Rana and K.K. Kim, "NAS-OD: Neural Architecture Search for Object Detection," *2024 International Conference on Electronics, Information, and Communication, ICEIC 2024*, 2024. DOI: https://doi.org/10.1109/ICEIC61013.2024.10457265

[25] L. Chen, Q. Ding, Q. Zou, Z. Chen and L. Li, "DenseLightNet: A Light-Weight Vehicle Detection Network for Autonomous Driving," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 12, pp. 10600–10609, 2020. DOI: https://doi.org/10.1109/TIE.2019.2962413

[26] H. Zhang, C. Lu, and E. Chen, "Obstacle detection: improved YOLOX-S based on swin transformer-tiny," *Optoelectronics Letters*, vol. 19, no. 11, pp. 698–704, 2023. DOI: https://doi.org/10.1007/s11801-023-3018-9

[27] M. Kang, W. Lee, K. Hwang and Y. Yoon, "Vision Transformer for Detecting Critical Situations and Extracting Functional Scenario for Automated Vehicle Safety Assessment," *Sustainability (Switzerland)*, vol. 14, no. 15, 2022. DOI: https://doi.org/10.3390/su14159680

[28] Y. Qin, D. He, Z. Jin, Y. Chen and S. Shan, "An Improved Deep Learning Algorithm for Obstacle Detection in Complex Rail Transit Environments," *IEEE Sensors Journal*, vol. 24, no. 3, pp. 4011–4022, 2024. DOI: https://doi.org/10.1109/JSEN.2023.3340688

[29] M.R. Abdurrahman, H. Al-Aziz, F.A. Zayn, M.A. Purnomo, and H.A. Santoso, "Development of Robot Feature for Stunting Analysis Using Long-Short Term Memory (LSTM) Algorithm," *Journal of Informatics and Web Engineering*, vol. 3, no. 3, pp. 164–175, 2024, DOI: https://doi.org/10.33093/jiwe.2024.3.3.10

[30] R. Fang and C. Cai, "Computer vision based obstacle detection and target tracking for autonomous vehicles," *MATEC Web of Conferences*, vol. 336, p. 07004, 2021. DOI: https://doi.org/10.1051/matecconf/202133607004

[31] A. Masoumian, D.G.F. Marei, S. Abdulwahab, J. Cristiano, D. Puig and H.A. Rashwan, "Absolute Distance Prediction Based on Deep Learning Object Detection and Monocular Depth Estimation Models," *Frontiers in Artificial Intelligence and Applications*, pp. 325–334, 2021. DOI: https://doi.org/10.3233/FAIA210151

[32] X. Chen, Y. Liu, and K. Achuthan, "WODIS: Water Obstacle Detection Network Based on Image Segmentation for Autonomous Surface Vehicles in Maritime Environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, 2021. DOI: https://doi.org/10.1109/TIM.2021.3092070

[33] H. Kim *et al.*, "Vision-Based Real-Time Obstacle Segmentation Algorithm for Autonomous Surface Vehicle," *IEEE Access*, vol. 7, pp. 179420–179428, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2959312

[34] P.S. Perumal *et al.*, "LaneScanNET: A deep-learning approach for simultaneous detection of obstacle-lane states for autonomous driving systems," *Expert Systems with Applications*, vol. 233, 2023. DOI: https://doi.org/10.1016/j.eswa.2023.120970

[35] Y. Xie *et al.*, "SparseFusion: Fusing Multi-Modal Sparse Representations for Multi-Sensor 3D Object Detection," *arXiv*, 2023. DOI: https://doi.org/10.48550/arXiv.2304.14340

[36] T. Liu, S. Du, C. Liang, B. Zhang, and R. Feng, "A Novel Multi-Sensor Fusion Based Object Detection and Recognition Algorithm for Intelligent Assisted Driving," *IEEE Access*, vol. 9, pp. 81564–81574, 2021. DOI: https://doi.org/10.1109/ACCESS.2021.3083503

[37] C. Zhang *et al.*, "Robust-FusionNet: Deep Multimodal Sensor Fusion for 3-D Object Detection Under Severe Weather Conditions," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, 2022. DOI: https://doi.org/10.1109/TIM.2022.3191724

[38] H. Xiang, R. Xu and J. Ma, "HM-ViT: Hetero-modal Vehicle-to-Vehicle Cooperative perception with vision transformer," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 284–295, 2023. DOI: https://doi.org/10.1109/ICCV51070.2023.00033

[39] Y. He *et al.*, "BiViT: Extremely Compressed Binary Vision Transformer," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5628–5640, Nov. 2022. DOI: https://doi.org/10.1109/ICCV51070.2023.00520.

[40] R. Xu, C.J. Chen, Z. Tu, and M.H. Yang, "V2X-ViTv2: Improved Vision Transformers for Vehicle-to-Everything Cooperative Perception," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 1, pp. 650–662, 2025. DOI: https://doi.org/10.1109/TPAMI.2024.3479222

[41] D. Minott, S. Siddiqui and R.J. Haddad, "Benchmarking Edge AI Platforms: Performance Analysis of NVIDIA Jetson and Raspberry Pi 5 with Coral TPU," *Conference Proceedings - IEEE SOUTHEASTCON*, pp. 1384–1389, 2025. DOI:https://doi.org/10.1109/SoutheastCon56624.2025.10971592

[42] S. Tennekoon, N. Wedasingha, A. Welhenge, N. Abhayasinghe and I. Murray Am, "Advancing Object Detection: A Narrative Review of Evolving Techniques and Their Navigation Applications," *IEEE Access*, vol. 13, pp. 50534–50555, 2025. DOI: https://doi.org/10.1109/ACCESS.2025.3551686

[43] I. Atik, "Deep Learning in Military Object Detection: An Example of the Yolo-Nas Model," *8th International Symposium on Innovative Approaches in Smart Technologies, ISAS 2024 - Proceedings*, 2024. DOI: https://doi.org/10.1109/ISAS64331.2024.10845459

[44] R. Varghese and M. Sambath, "A Comprehensive Review On Two-Stage Object Detection Algorithms," *iQ-CCHESS 2023 - 2023 IEEE International Conference on Quantum Technologies, Communications, Computing, Hardware and Embedded Systems Security*, 2023. DOI:https://doi.org/10.1109/IQCCHESS56596.2023.10391506

[45] J.G. Min, D. Kam, Y. Byun, G. Park and Y. Lee, "Energy-Efficient RISC-V-Based Vector Processor for Cache-Aware Structurally-Pruned Transformers," *Proceedings of the International Symposium on Low Power Electronics and Design*, vol. 2023-August, 2023. DOI:https://doi.org/10.1109/ISLPED58423.2023.10244508

[46] Z. Ouardirhi, S.A. Mahmoudi and M. Zbakh, "Enhancing Object Detection in Smart Video Surveillance: A Survey of Occlusion-Handling Approaches," *Electronics 2024, Vol. 13, Page 541*, vol. 13, no. 3, p. 541, 2024. DOI: https://doi.org/10.3390/ELECTRONICS13030541

[47] Y. Chen and W. Zhou, "Hybrid-Attention Network for RGB-D Salient Object Detection," *Applied Sciences 2020, Vol. 10, Applied Sciences*, vol. 10, no. 17, p. 5806, 2020. DOI: https://doi.org/10.3390/APP10175806

[48] Z. Chen, Z. Ding, X. Zhang, X. Zhang and T. Qin, "Improving Out-of-Distribution Generalization in SAR Image Scene Classification with Limited Training Samples," *Remote Sensing 2023,* vol. 15, no. 24, p. 5761, 2023. DOI: https://doi.org/10.3390/RS15245761

[49] H. Mousazadeh *et al.*, "Ships and Offshore Structures Dynamic and static object detection and tracking in an autonomous surface vehicle Dynamic and static object detection and tracking in an autonomous surface vehicle," *Ocean Engineering*, vol. 159, pp. 56-65, 2018. DOI: https://doi.org/10.1016/j.oceaneng.2018.04.018

[50] S. Feng, B. Sebastian and P. Ben-Tzvi, "A collision avoidance method based on deep reinforcement learning," *Robotics*, vol. 10, no. 2, 2021. DOI: https://doi.org/10.3390/robotics10020073