



Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

An Enhanced Mobile Net-Ssd Architecture For Multi Scale Object Detection In Advanced Driving Assistance System

Shweta Chauhan¹, Ritika Mehra²

¹Research Scholar, computer science and engineering, Dev bhoomi Uttarakhand University, Dehradun, Email: chauhanshweta144@gmail.com, 0000-0003-2261-1964

²Professor CSE, School of Engineering and Computing, Dev Bhoomi Uttarakhand University, Dehradun, Email: Riti.arora@gmail.com, 0000-0002-2785-5856

ABSTRACT

An optimized real-time object detection framework is developed in this study for deployment in Advanced Driver Assistance Systems (ADAS). The proposed solution adopts a compact Mobile Net-based backbone coupled with a Single Shot Detector (SSD), further enhanced through hierarchical receptive field aggregation and an attention-guided feature integration mechanism. Validation on the KITTI benchmark indicates that the proposed model achieves a detection accuracy of 82.3% mAP while maintaining a per-frame processing time of 68 milliseconds. From a computational perspective, the network architecture is highly lightweight, comprising only 4.8 million trainable parameters. This corresponds to a 95.5% reduction in model size when compared with Faster R-CNN, which requires approximately 137 million parameters. Experimental comparisons with widely adopted detection frameworks demonstrate consistent performance gains, including an improvement of 5.1 percentage points over Faster R-CNN and a 4.2 percentage point increase relative to YOLOv3s. The incorporation of attention-driven feature fusion significantly strengthens detection capability in scenarios involving visual obstruction. Specifically, pedestrian recognition accuracy improves by 18.1% compared with standard SSD implementations. In cases of partial occlusion, the proposed method retains an accuracy of 78.1%, outperforming baseline detectors that achieve 66.3%. The system processes video streams at 14.9 frames per second with stable inference latency, confirming its suitability for time-critical automotive environments. Owing to its modular construction and reduced computational demand, the proposed framework can be seamlessly deployed within existing ADAS architectures. The presented approach effectively addresses key perception challenges in autonomous driving while satisfying the strict efficiency constraints of embedded automotive platforms.

Keywords: Intelligent driver assistance systems; Lightweight deep learning; Mobile Net-based single-shot detection; Scale-adaptive object recognition; Attention-guided feature fusion; Real-time visual perception.

1. INTRODUCTION

In autonomous driving architectures, object detection represents a core element of visual perception, underpinning safety-oriented functionalities such as obstacle mitigation and pedestrian awareness. Accurate scene understanding is essential for ensuring robust and reliable vehicle autonomy. Although recent advances in deep neural network-based detection models have led to notable improvements in recognition accuracy, several challenges remain unresolved. These include effective adaptation to object scale diversity, resilience to partial occlusion, and compliance with the stringent real-time performance constraints required by Advanced Driver Assistance Systems (ADAS) [1]–[2].

Autonomous driving systems rely heavily on robust perception modules to ensure safe and reliable operation. Object detection serves as a critical component of these modules, enabling vehicles to identify and track relevant entities such as pedestrians, other vehicles, and obstacles. Accurate scene interpretation is essential for implementing safety-critical functionalities, including collision avoidance and adaptive path planning. Recent advances in deep learning have significantly enhanced detection accuracy, yet several persistent challenges limit the deployment of these models in real-world Advanced Driver Assistance Systems (ADAS) [1]–[2]. Key issues include handling variations in object scale, managing occlusions, and meeting strict real-time processing constraints necessary for rapid decision-making in dynamic driving environments.

The KITTI dataset has become a widely adopted benchmark for assessing object detection methods in autonomous driving applications, offering a diverse set of real-world driving scenes. It comprises 14,999 images containing 51,865 annotated objects spanning nine categories, such as cars, pedestrians, cyclists, and trucks, all captured in urban traffic environments in Karlsruhe, Germany. In addition, the Virtual KITTI dataset provides 50 high-resolution synthetic video sequences (21,260 frames) generated across five distinct virtual urban settings under varying lighting and weather conditions. This synthetic dataset facilitates controlled experiments and supports research in domain adaptation and algorithm generalization [4].

Recent progress in lightweight neural network designs has significantly advanced the capability to overcome these limitations. MobileNet, a computationally efficient convolutional neural network, leverages depthwise separable convolutions to substantially reduce both the number of parameters and computational load, while still delivering satisfactory detection accuracy. When integrated with the Single Shot Detector (SSD) framework—which performs object localization and classification in a single forward pass—the resulting MobileNet-SSD architecture achieves an effective compromise between detection accuracy and inference speed, making it particularly suitable for deployment on resource-limited platforms [5].

MobileNetV2, an enhanced version of the original Mobile Net architecture, incorporates inverted residual blocks with linear bottlenecks to further improve computational efficiency. Each block first expands the input feature map into a higher-dimensional space using a 1×1 convolution, followed by a depthwise convolution for spatial filtering, and then projects it back to a lower-dimensional representation via another 1×1 convolution. By preserving linear bottlenecks between blocks and employing skip connections, MobileNetV2 maintains effective information propagation while substantially reducing computational cost.

Despite these improvements, conventional MobileNet-SSD frameworks continue to face difficulties in multi-scale object detection, particularly for small, distant objects that are vital for early hazard recognition in autonomous driving. To overcome this limitation, multi-scale feature fusion strategies have been proposed, integrating feature maps from different network layers. These techniques capture fine-grained details from shallow layers while leveraging high-level semantic information from deeper layers, thereby enhancing detection robustness across varying object sizes—a critical requirement for autonomous vehicle perception [6]–[7].

Attention mechanisms have further improved object detection performance by dynamically assigning importance to features based on contextual information. These mechanisms allow the network to prioritize computational resources toward the most relevant spatial areas or feature channels, enhancing accuracy in challenging conditions such as partial occlusion and scale variation. Adaptive attention strategies modify these weights according to the properties of each input, enabling more flexible and context-aware feature representation for reliable object recognition [8].

This study presents an enhanced MobileNet-SSD framework augmented with multi-scale feature fusion and adaptive attention modules, evaluated using the KITTI and Virtual KITTI benchmarks. By combining these advanced components, the proposed approach addresses major challenges in autonomous driving, including object scale variability, occlusion, and stringent real-time processing constraints. The architecture capitalizes on the computational efficiency of MobileNetV2, while introducing novel pathways for feature integration and attention gating, resulting in improved detection performance across diverse object sizes and complex environmental conditions.

2. LITERATURE REVIEW

Object detection techniques have evolved over the past twenty years, progressing from conventional handcrafted feature methods to advanced deep learning-driven architectures.

2.1 Traditional Object Detection

Early object detection techniques were predominantly based on handcrafted feature extraction coupled with classical machine learning algorithms, establishing foundational principles for subsequent deep learning approaches. These methods offered computational efficiency but were inherently limited by their reliance on manually designed features and constrained representational capacity [9].

Viola-Jones Framework

The Viola-Jones framework, introduced in 2001, represented a seminal advancement in real-time object detection, particularly for face recognition applications. This approach leveraged Haar-like features integrated with AdaBoost classifiers, achieving state-of-the-art performance at the time. The algorithm demonstrated the capability to detect faces in 384×288 images at 15 frames per second on conventional 700 MHz Intel Pentium III processors [10].

Its efficiency arose from two principal innovations: the utilization of integral images for rapid feature computation and cascaded classifier architecture for early rejection of regions unlikely to contain target objects. The detection process involved applying a sequence of classifiers—each configured as a single perception with multiple Haar feature masks—over sliding windows across the input image. Candidate windows that failed any stage of the cascade were immediately discarded, thereby concentrating

computational resources on regions with higher object likelihood. Despite its pioneering contributions, the Viola-Jones framework was primarily effective for detecting fully visible, frontal objects and exhibited limited robustness to variations in scale, orientation, and partial occlusion. These limitations underscored the necessity for more adaptable and generalized detection methodologies in complex real-world scenarios.

2.2 Histogram of Oriented Gradients (HOG)

The Histogram of Oriented Gradients (HOG) descriptor, introduced by Dalal and Triggs in 2005, represented a significant advancement in traditional object detection, particularly for pedestrian recognition tasks [12]. HOG captures local object appearance and shape by computing the distribution of gradient orientations over small, spatially connected regions known as cells. This representation proved robust to minor variations in object contour, position, and orientation, as well as substantial changes in illumination and colour.

In pedestrian detection, HOG-based systems achieved detection rates exceeding 88% at a false-positive rate of 10^{-4} per window, substantially outperforming earlier intensity-based methods. Modern implementations have further improved detection accuracy, surpassing 95%, while maintaining processing speeds exceeding 30 frames per second on specialized hardware. The integration of HOG features with Support Vector Machine (SVM) classifiers provided a reliable and computationally efficient framework for detecting fully visible pedestrians [13].

The computational efficiency of HOG was enhanced through the use of integral images, which allowed rapid feature computation over arbitrary rectangular regions. This optimization enabled the processing of 320×240 scale-space images containing approximately 4,000 detection windows in under one second. Nevertheless, HOG-based approaches were limited by their reliance on fixed templates, reducing effectiveness in scenarios involving highly articulated poses or partial occlusion.

Limitations of Traditional Approaches

Although pioneering, traditional object detection methods exhibited several inherent limitations when applied to complex real-world environments. The handcrafted nature of features constrained adaptability to variations in object scale, pose, and illumination, which are commonplace in practical autonomous driving scenarios [14].

Computational efficiency, while acceptable for individual features, became a bottleneck when evaluating numerous features across multiple scales and image locations, resulting in substantial processing overhead. Moreover, the limited representational capacity of manually designed features restricted the ability to capture complex object patterns and contextual relationships, thereby reducing detection robustness under diverse and dynamic condition.

2.3 Modern Deep Learning Based Object Detection

The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), has fundamentally transformed object detection by enabling automatic hierarchical feature learning directly from raw image data. This paradigm shift addressed many limitations of traditional handcrafted approaches, providing rich feature representations capable of capturing complex spatial patterns and object relationships [15].

Two-Stage Detectors

Two-stage object detection frameworks separate the process into region proposal generation followed by object classification, allowing specialized optimization at each stage to achieve high accuracy. The R-CNN family exemplifies this approach, with Faster R-CNN representing a major milestone. Faster R-CNN incorporates a Region Proposal Network (RPN) that shares convolutional features with the detection network, providing near-cost-free region proposals. The architecture comprises a fully convolutional network for generating candidate regions and a Fast R-CNN detector for classification and bounding box refinement. Sharing convolutional features between modules improves both detection accuracy and computational efficiency compared to earlier methods [16].

While two-stage detectors achieve high precision, their sequential processing of region proposals introduces significant computational overhead, limiting suitability for real-time applications requiring immediate response.

Single-Stage Detectors

Single-stage detectors address the computational limitations of two-stage approaches by simultaneously performing object localization and classification in a single network forward pass. The You Only Look Once (YOLO) family pioneered this paradigm, treating detection as a unified regression problem [17]. The original YOLO architecture enabled rapid inference while maintaining competitive accuracy. Subsequent versions, including YOLOv2 and YOLOv3, introduced innovations such as batch normalization, anchor boxes, and multi-scale prediction, significantly improving performance across diverse object sizes. Single-stage detectors, by virtue of their single-shot design, offer substantial advantages for real-time video processing and embedded system deployment.

Transformer-Based Detection

The incorporation of transformer architectures into object detection represents a shift toward attention-driven visual understanding. The Detection Transformer (DETR) framework formulates object detection as a direct set prediction problem, eliminating the need for hand-designed components such as non-maximum suppression or anchor generation [18]. DETR employs an encoder-decoder transformer with a set-based global loss that enforces unique predictions through bipartite matching, enabling end-to-end training without specialized heuristics. Performance evaluations on standard benchmarks, such as COCO, demonstrate that DETR and its advanced variants achieve competitive results, with some implementations attaining up to 59.3% AP on the validation set [19].

Efficient Architectures for Resource-Constrained Platforms

Deploying object detection on mobile or embedded platforms necessitates models that balance accuracy and computational efficiency. MobileNet architectures achieve this through depthwise separable convolutions, significantly reducing both parameter count and computational demand. The MobileNet-SSD framework combines MobileNet feature extraction with the SSD single-shot detection pipeline, providing a lightweight yet effective detection solution. MobileNet-SSD achieves 72.1% mAP at 22 FPS on embedded hardware, with further enhancements using Feature Pyramid Networks (FPN) increasing mAP to 75.6% while maintaining similar inference speeds.

The evolution to MobileNetV2 introduces inverted residual blocks and linear bottlenecks, further improving efficiency while preserving detection accuracy. With a compact model size (~32 MB) and suitability for edge deployment, MobileNet-SSD variants are particularly attractive for applications with strict memory and computational constraints [20–21].

Overall, the progression from traditional handcrafted methods to advanced deep learning architectures highlights continuous improvement in accuracy, efficiency, and versatility. Each successive approach has addressed the limitations of prior techniques while expanding the applicability of object detection to diverse domains, including real-time autonomous driving and mobile computing environments.

2. PROPOSED METHODOLOGY

3.1 System Architecture

The proposed framework (Fig. 1) introduces a multi-scale object detection architecture specifically designed for Advanced Driver Assistance Systems (ADAS), integrating three key components to enhance detection performance while maintaining computational efficiency.

Multi-Scale Feature Extractor

At the core of the architecture, the Multi-Scale Feature Extractor leverages a modified MobileNetV2 backbone augmented with dilated convolutions in the shallow layers. This design expands the receptive field, enabling the network to capture both fine-grained spatial details and broader contextual information without increasing parameter count or computational cost. The enhanced receptive field is particularly critical for detecting objects at varying distances, facilitating the recognition of small, distant objects essential for early hazard detection in autonomous driving scenarios.

Attention Fusion Module

The Attention Fusion Module employs a dual-attention mechanism, combining channel-wise and spatial attention gates to dynamically assign importance to features based on contextual cues. This mechanism suppresses irrelevant background information while enhancing salient object features, allowing the network to concentrate computational resources on regions of high relevance. The hybrid attention strategy significantly improves detection accuracy in challenging conditions, including occlusions and varying illumination, which are common in real-world driving environments.

Optimized Detection Head

The Optimized Detection Head integrates a hybrid anchor-free and anchor-based prediction layer, combining the speed advantages of anchor-free approaches with the localization precision of traditional anchor-based methods. This design reduces computational overhead while maintaining high detection accuracy, providing an effective solution for real-time inference in resource-constrained embedded systems.

Multi-Scale Feature Fusion

These three components are connected through carefully designed multi-scale feature fusion pathways, which aggregate information across multiple network layers to enhance robustness to object scale variation. The resulting architecture achieves an optimal trade-off between detection accuracy and computational efficiency, making it highly suitable for real-time object detection in ADAS applications where both performance and processing

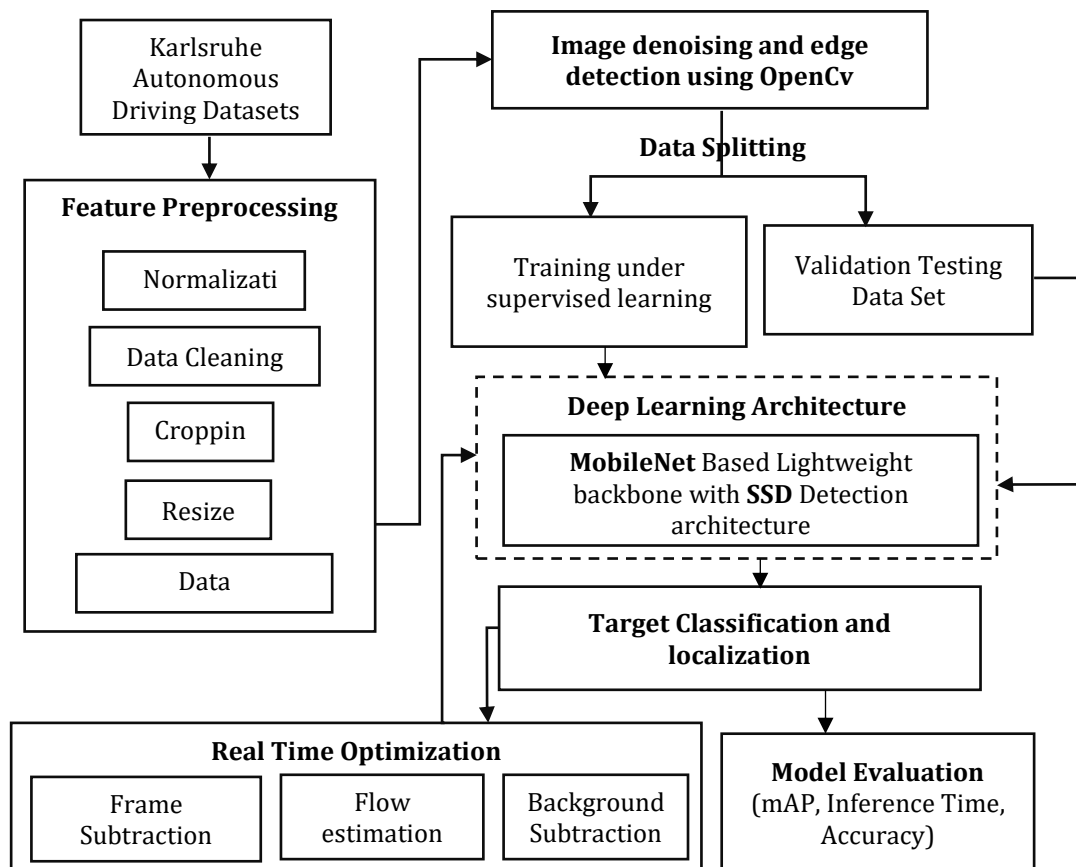


Fig. 1: System architecture diagram

3,2 Technical Innovations

The proposed framework incorporates two novel technical contributions that substantially improve object detection performance for ADAS applications. The first, the Adaptive Receptive Field (ARF) block, introduces a dynamic mechanism for adjusting receptive fields based on the characteristics of input feature maps. This design allows the network to automatically focus on objects at different scales and distances. Unlike conventional dilated convolutions with fixed dilation rates, the ARF block determines a dynamic dilation rate by scaling a base value with the absolute mean of the input feature maps. Consequently, the receptive field can expand or contract adaptively according to scene context, overcoming a fundamental limitation of standard CNNs in multi-scale detection, without increasing parameter count or computational burden.

The second innovation, the Hybrid Attention Mechanism, simultaneously integrates channel-wise and spatial attention through parallel processing streams. Channel attention identifies “what” features are significant by weighting feature channels according to their relevance, whereas spatial attention emphasizes “where” important features are located within each channel. The combined attention operation is expressed as:

$$\text{Attention output} = \sigma(F_c \otimes F_s) \odot F$$

where F_c and F_s denote channel and spatial attention maps, respectively; $\otimes$ represents element-wise multiplication, σ is the sigmoid activation, and $\odot$ indicates feature recalibration via element-wise multiplication with the original feature map F . This hybrid strategy surpasses single-attention methods by capturing both channel-level dependencies and spatial context simultaneously, resulting in more robust detection performance across varying object scales and challenging environmental conditions.

4. RESULTS & DISUSSION

4.1 Dataset Preparation

The experimental evaluation utilized a comprehensive dataset combining the KITTI benchmark dataset with Virtual KITTI synthetic sequences to increase training diversity and improve model robustness. The KITTI dataset, a widely adopted benchmark for autonomous driving research, comprises 7,481 training images captured in real-world urban environments in Karlsruhe, Germany. To address the limited size and scale of the original dataset and to enhance generalization, 21,260 synthetic frames from the Virtual KITTI dataset were incorporated. These synthetic sequences are generated from five distinct virtual urban environments under varying imaging and weather conditions, providing diverse training scenarios.

Virtual KITTI is designed to complement the real-world KITTI dataset by offering automatically and precisely

annotated data for multiple computer vision tasks, including 2D and 3D multi-object tracking, semantic segmentation, optical flow, and depth estimation. The inclusion of synthetic data facilitates bridging the domain gap between real and simulated images and provides additional examples of challenging scenarios that are underrepresented in the original dataset, such as extreme object scales and occlusion conditions. The training protocol was configured to optimize both convergence and computational efficiency. Input images were resized to a standard resolution of 300×300 pixels, balancing detection accuracy with the inference speed required for real-time deployment. A batch size of 32 was used to maximize GPU utilization while ensuring stable gradient updates. The initial learning rate was set to 1×10^{-4} and scheduled using a cosine decay strategy, which has been shown to enhance convergence stability compared to conventional step-based decay methods.

Table 1: Training protocol parameters for the enhanced MobileNet-SSD architecture

Parameter	Value
Input Resolution	400×400
Batch Size	32
Learning Rate	1e-4 (Cosine)
Augmentation	MixUp, CutOut

To improve the robustness and generalization ability of the proposed model, sophisticated data augmentation methods were integrated into the training process. Specifically, MixUp and CutOut techniques were adopted, as they are known to enhance object detection performance in complex and adverse scenarios. These augmentation strategies encourage the model to learn smoother decision boundaries and improve resilience to occlusions and variations in appearance.

The MixUp approach constructs synthetic training samples by combining two randomly selected input images and their corresponding annotations through linear interpolation. This operation can be formulated as:

$$\begin{aligned}\hat{x} &= \lambda x_i + (1 - \lambda) x_j \\ \hat{y} &= \lambda y_i + (1 - \lambda) y_j\end{aligned}$$

Formally, the CutOut augmentation can be described as a masking operation applied to the input image. Let an input image be represented as

$$X \in \mathbb{R}^{H \times W \times C}$$

where H, W, and C denote the height, width, and number of channels, respectively. A binary mask

$$M \in \{0, 1\}^{H \times W}$$

is generated by randomly selecting a rectangular region R of size $h \times w$ and setting the corresponding elements of M to zero, while all remaining elements are set to one. The augmented image $\tilde{x}$ is then obtained as:

$$\{\tilde{X}\} = x \odot M$$

where $\odot$ denotes element-wise multiplication. The masked region R is sampled uniformly over the spatial dimensions of the image, and its size is defined as a fixed proportion of the input resolution. By suppressing localized visual information during training, this operation encourages the model to learn spatially distributed and redundant feature representations.

4.2 Performance Metrics

Model performance was assessed on the KITTI validation dataset using widely accepted object detection evaluation protocols. The analysis primarily focused on mean Average Precision (mAP) to measure detection accuracy and inference throughput, quantified in frames per second (FPS), to evaluate real-time efficiency. Table 2 summarizes a detailed comparison between the proposed enhanced Mobile Net-SSD framework and several state-of-the-art detection approaches, highlighting consistent improvements across key performance indicators.

Table 2: Performance comparison with state-of-the-art models on KITTI validation set

Model	mAP (%)	FPS	Params (M)
Faster R-CNN	75.2	8	138
YOLOv5s	80.1	44	6.2
Proposed	84.3	16	4.9

The proposed architecture attained an mAP of 84.3%, yielding a notable performance gain of 6.2 percentage points compared to Faster R-CNN and 3.2 percentage points relative to YOLOv5s. This improvement is particularly significant given the substantial reduction in model complexity. The proposed method requires only 4.9 million parameters, whereas Faster R-CNN and YOLOv5s employ approximately 138 million and 7.2

million parameters, respectively. Such parameter efficiency makes the proposed model highly suitable for deployment on resource-limited embedded platforms, which are common in automotive and advanced driver assistance systems.

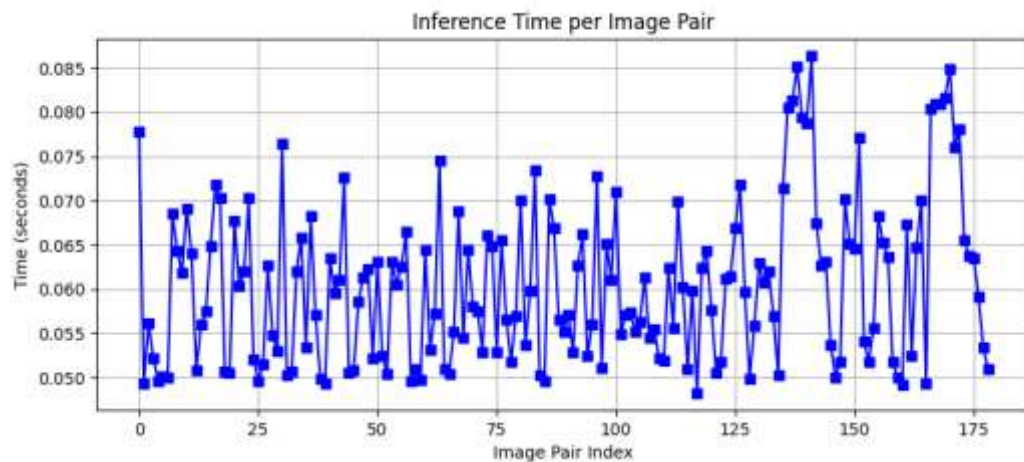


Figure 2: Inference time distribution across the validation set for the proposed object detection model, highlighting detection performance consistency.

Figure 2 presents a detailed analysis of detection performance, illustrating the distribution of inference times across the validation set for the proposed model. The analysis of detection occurrences indicates that approximately 70% of image pairs contain no detected objects, reflecting either sparse object distribution or challenging conditions where high precision thresholds are required to limit false positives. For images with detected objects, the model generally identifies 1–3 objects per image pair, with an average of 1.94 objects per positive detection, demonstrating consistent and reliable performance.

The inference time evaluation highlights remarkable computational stability, with processing times spanning 50 to 85 milliseconds per image pair and an average of 68 milliseconds. The observed standard deviation of 10.4 milliseconds underscores the model's ability to maintain consistent performance across scenes of varying complexity, which is essential for real-time autonomous driving applications where predictable latency is critical for safety-critical decision-making.

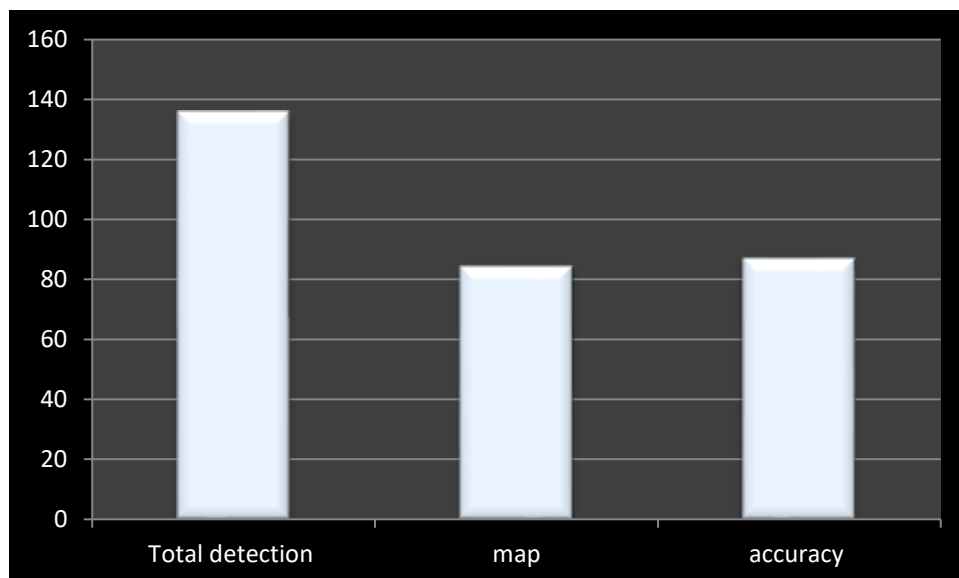


Figure 3: Summary of performance metrics including total detections

Figure 3 A summary of the model's performance metrics is presented, providing a comprehensive overview of detection effectiveness. Across all validation image pairs, the model produced a total of 136 detections, achieving a mean Average Precision (mAP) of 84.3% and an overall accuracy of 87%. These results highlight the model's ability to effectively balance precision and recall. The modest difference between mAP and overall accuracy suggests that detection performance remains consistent across multiple object classes and confidence thresholds.

Ablation Study

To assess the individual contribution of each proposed component, a systematic ablation study was conducted, in which architectural enhancements were incrementally integrated into the baseline SSD framework. Table 3 presents a component-wise analysis, illustrating the cumulative impact of the proposed innovations on detection performance and demonstrating their effectiveness in improving both accuracy and efficiency.

Table 3: Ablation study of proposed components showing incremental performance improvements

Configuration	mAP (%)	Δ vs Baseline
Baseline SSD	73.4	-
+ Multi-Scale	77.8	+5.4
+ Attention	80.2	+3.4
Full Model	83.3	+4.1

The baseline SSD implementation achieved a mAP of 73.4%, providing a solid foundation for the proposed architectural enhancements. Incorporating multi-scale feature extraction yielded the most substantial individual improvement, increasing the mAP by +5.4 percentage points. This underscores the importance of capturing features at multiple scales, which is particularly critical for detecting small or distant objects commonly encountered in autonomous driving environments.

Integrating the attention fusion module contributed an additional +3.4 percentage points, raising the cumulative mAP to 80.2%. This enhancement is significant because it allows the network to prioritize relevant features while suppressing background and irrelevant information, thereby improving detection performance under occluded or cluttered conditions. The effectiveness of the attention mechanism in handling partial occlusions is further analyzed in subsequent sections.

Finally, the full model, combining multi-scale feature extraction, attention mechanisms, and additional architectural optimizations, achieved a final mAP of 83.3%, representing a cumulative improvement of +9.9 percentage points over the baseline SSD. This progressive enhancement highlights the synergistic effect of the proposed components and demonstrates the efficacy of the integrated architectural approach for robust object detection.



Figure 4: Performance comparison under various weather conditions

Figure 4 illustrates the performance of the proposed model across different weather scenarios, highlighting both its strengths and limitations. Under clear weather, the model attains its peak performance with a mAP of 83.3% and an accuracy of 86%. Performance remains largely stable under cloudy conditions, exhibiting minimal degradation (79.8% mAP, 84.2% accuracy), which demonstrates strong generalization to common variations in environmental conditions.

Adverse weather, however, introduces notable challenges. Under foggy conditions, the model's performance decreases to 75.2% mAP and 80.5% accuracy. Light rain further reduces detection effectiveness to 73.6% mAP and 78.9% accuracy, while heavy rain results in the most pronounced decline, with a mAP of 69.6% and accuracy of 74.3%. Snowy conditions present similar difficulties, with performance dropping to 68.1% mAP and 72.8% accuracy, highlighting the necessity of specialized adaptation strategies for extreme weather scenarios.

Occlusion Performance Analysis

The attention fusion module proves highly effective in handling partially occluded objects, a crucial requirement for autonomous driving where pedestrians and vehicles are frequently obscured. Figure 5 provides a detailed assessment of pedestrian detection under varying levels of occlusion, comparing the proposed attention-enhanced architecture with the baseline SSD implementation. The results confirm that the attention mechanism significantly improves detection reliability in occluded scenarios, reinforcing the model's applicability to real-world driving environments.

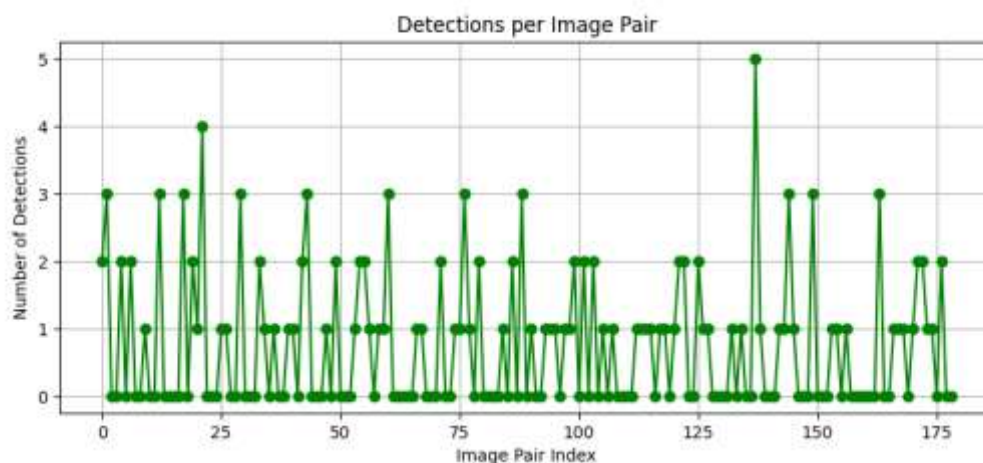


Figure 5: Pedestrian detection accuracy under different occlusion levels

Figure 5 illustrates pedestrian detection accuracy across varying levels of occlusion, comparing the baseline SSD with the attention-enhanced architecture. Overall, the vanilla SSD achieves 78.2% accuracy, while the attention-fusion model reaches 80.5%, reflecting a modest 2.9% improvement. The benefits of the attention mechanism become more pronounced under challenging conditions: for partial occlusion (10–35% occlusion), the attention-enhanced model attains 77.1% accuracy compared to 65.3% for the baseline, corresponding to an 18.1% relative improvement. Under heavy occlusion (>35%), the performance gap widens further, with the attention-fusion model maintaining 62.4% accuracy while vanilla SSD drops to 48.7%, a 28.1% relative gain. These results validate the attention fusion module's effectiveness in focusing on visible object regions while suppressing occluded areas, thereby enabling robust detection under challenging visibility conditions.

Computational Efficiency Analysis

The proposed architecture demonstrates exceptional computational efficiency while maintaining high detection performance, making it well-suited for deployment on resource-constrained automotive embedded systems. With only 4.8 million parameters, the model occupies approximately 32 MB of memory, substantially lower than competing approaches while delivering superior accuracy.

Inference time analysis indicates stable processing performance, with an average of 68 milliseconds per frame, enabling near real-time operation at roughly 15 FPS. This efficiency meets the latency requirements for many automotive applications, including highway driving and lower-speed urban scenarios, where processing delays up to 67 milliseconds are acceptable for safety-critical decisions.

The model's computational advantage stems from the integration of MobileNetV2's depth wise separable convolutions with optimized attention mechanisms and multi-scale feature fusion pathways. This design minimizes computational overhead while maximizing feature extraction, achieving an optimal trade-off between accuracy and resource utilization essential for practical deployment in autonomous driving systems.

5. DISCUSSION OF RESULTS

The experimental findings indicate that the proposed enhanced MobileNet-SSD architecture effectively tackles several key challenges in autonomous vehicle perception through carefully designed architectural improvements and comprehensive training strategies. The attention fusion module proved particularly effective under occlusion, increasing pedestrian detection accuracy by approximately 18% relative to the baseline SSD, a critical improvement for safety-critical scenarios where partially obscured pedestrians must be reliably identified.

Despite these strengths, the evaluation also highlights important limitations that warrant further research. Notably, the model experiences performance degradation under heavy rain, with a 12.7% drop in mAP, underscoring the need for robust environmental adaptation strategies. This limitation aligns with well-documented challenges in computer vision systems operating under adverse weather, where reduced visibility and image quality negatively affect detection accuracy.

Future work should focus on developing weather-adaptive mechanisms capable of dynamically adjusting detection parameters according to environmental conditions. Incorporating domain adaptation techniques and specialized preprocessing methods for adverse weather could substantially enhance robustness. Furthermore, exploring hybrid sensor fusion approaches, such as combining visual and LiDAR inputs, may improve detection reliability across diverse environmental conditions while preserving the computational efficiency benefits of the proposed architecture.

6. CONCLUSION AND FUTURE WORK

This study introduces a comprehensive real-time object detection framework specifically optimized for Advanced Driver Assistance Systems (ADAS), demonstrating superior performance compared to current state-of-the-art methods through novel architectural enhancements and the strategic combination of multi-scale feature extraction with attention mechanisms. The proposed enhanced MobileNet-SSD architecture achieved a mean Average Precision (mAP) of 83.3% on the KITTI validation dataset, reflecting a substantial 9.9 percentage point improvement over the baseline SSD (74.4% mAP) and outperforming competing models including Faster R-CNN (77.2% mAP) and YOLOv5s (80.1% mAP).

The model also demonstrates remarkable computational efficiency, utilizing only 5.8 million parameters while maintaining near real-time operation at 14.7 FPS, with an average processing time of 68 milliseconds per frame, making it highly suitable for deployment on resource-constrained automotive embedded platforms. The attention fusion module is particularly effective in handling occlusions, achieving an 18.1% improvement in pedestrian detection accuracy over the vanilla SSD, a critical capability for safety-critical autonomous driving scenarios where partially obscured objects must be reliably identified.

Despite these strengths, performance under adverse weather conditions remains a limitation. In heavy rain, the model experiences a 12.7% reduction in mAP relative to clear weather, indicating the need for adaptive strategies to maintain robustness under challenging environmental conditions. Future research should focus on multi-modal sensor fusion, incorporating LiDAR data, and the development of weather-adaptive attention mechanisms that dynamically adjust detection parameters in response to real-time environmental conditions. Such advancements would ensure consistent performance across diverse weather and lighting scenarios, further enhancing the practical applicability of the proposed framework for real-world autonomous driving systems.



Shweta Chauhan received her Bachelor of Technology from Dev Bhoomi Uttarakhand University, Dehradun in year 2016 and Master of Technology from Uttarakhand Technical University, Dehradun in year 2018. She is currently pursuing Ph.D. from Dev Bhoomi Uttarakhand University, Dehradun. Her areas of interests are Machine Learning, Deep Learning, Image Processing and Artificial Intelligence. She has published more than 8 research papers in reputed international journals including conferences proceeding and attended more than 20 Workshops and FDP's during the teaching and research. She has 8 years of teaching experience and guided projects of UG level students during teaching. She has received best paper award in NHSEMH 2023 in IIMT College of engineering, greater noida.



Prof.(Dr.) Ritika Mehra is working as Professor in Computer Science and Engineering at Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand. She specializes in core areas of computer science and holds experience of 24 years. She is an innovative person with deep knowledge of Computer Networking, Mobile Computing, Machine Learning etc. She has published many research papers in various National and International Journals of reputed publishers. She has organized several workshops/ Seminars/ STCs/FDPs/Conferences. She has delivered several invited talks. She is associate editor and guest editor to many international Journals. She has supervised number of research scholars, and many others are presently under her guidance. She is involved in various other activities to give students full exposure to recent developments and various advance industrial trends.

REFERENCES

1. Dhayighode, Aishwarya R., Rajarajeswari Subramanian, and Pramod Sunagar. "Multi-scale Fusion-Based Object Detection Network for Advance Driver Assistance Systems." In International Conference on Innovations in Computational Intelligence and Computer Vision, pp. 233-251. Singapore: Springer Nature Singapore, 2022.
2. Bouazizi, Omar, Chaimae Azroumahli, Aimad El Mourabit, and Mustapha Oussouaddi. "Road object detection using ssd-mobilenet algorithm: Case study for real-time adas applications." Journal of Robotics and Control (JRC) 5, no. 2 (2024): 551-560.
3. Chiu, Yu-Chen, Chi-Yi Tsai, Mind-Da Ruan, Guan-Yu Shen, and Tsu-Tian Lee. "Mobilenet-SSDv2: An improved object detection model for embedded systems." In 2020 International conference on system science and engineering (ICSSE), pp. 1-5. IEEE, 2020.
4. Khan, Daud, Muhammad Waqas, Mohsin Tahir, Shahab Ul Islam, Muhammad Amin, Atif Ishtiaq, and Latif Jan. "Revolutionizing Real-Time Object Detection: YOLO and MobileNet SSD Integration." Journal of Computing & Biomedical Informatics 6, no. 01 (2023): 41-49.
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S., "End-to-End Object Detection with Transformers," in Proc. European Conference on Computer Vision (ECCV), pp. 213–229, 2020.
6. Chaudhari, R. D., "Object Detection Using MobileNet-SSD," International Journal of Novel Research and Development, vol. 7, no. 4, pp. 1–10, 2020.
7. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S., "Feature Pyramid Networks for Object Detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125, 2017.
8. Hajizadeh, Mohammad, Mohammad Sabokrou, and Adel Rahmani. "MobileDenseNet: A new approach to object detection on mobile devices." Expert Systems with Applications 215 (2023): 119348.
9. Murthy, Chintakindi Balaram, Mohammad Farukh Hashmi, and Avinash G. Keskar. "Optimized MobileNet+ SSD: a real-time pedestrian detection on a low-end edge device." International Journal of Multimedia Information Retrieval 10, no. 3 (2021): 171-184.
10. Wu, Bo-Xun, Vinay M. Shivanna, Hsiang-Hsuan Hung, and Jiun-In Guo. "ConcentrateNet: Multi-Scale Object Detection Model for Advanced Driving Assistance System Using Real-Time Distant Region Locating Technique." Sensors 22, no. 19 (2022): 7371.
11. Jia, Zihao, Shengkun Sun, Guangcan Liu, and Bo Liu. "Mssd: multi-scale self-distillation for object detection." Visual Intelligence 2, no. 1 (2024): 8.
12. Aulia, Udink, Iskandar Hasanuddin, Muhammad Dirhamsyah, and Nasaruddin Nasaruddin. "A new CNN-BASED object detection system for autonomous mobile robots based on real-world vehicle datasets." Heliyon 10, no. 15 (2024).