Lightweight Deep Learning Models for Real-Time Object Detection on Edge Devices: A Comparative Performance Study

Dr. Sandeep Kumar
Tula's Institute, Dehradun, U.K., India
Email: sandeep.kumar.cse@tulas.edu.in

Abstract

Real-time object detection at the network edge — on cameras, drones, vehicles, and embedded industrial controllers — has become a defining workload of applied computer vision. The governing constraint is no longer accuracy alone but the joint budget of latency, memory, energy, and thermal headroom available on resource-constrained hardware. This paper presents a structured comparative study of the principal lightweight detection approaches, organised around two axes: architectural efficiency strategies (compact backbone design, single-stage detection heads, multi-scale feature aggregation) and post-design compression strategies (quantisation, pruning, and knowledge distillation). We develop a comparison framework covering seven evaluation dimensions — detection paradigm, backbone design strategy, relative model footprint, latency behaviour, accuracy–efficiency trade-off profile, hardware affinity, and deployment tooling maturity — and apply it qualitatively across representative model families including the MobileNet-SSD lineage, the compact YOLO lineage, and compound-scaled detectors of the EfficientDet type. Because published benchmark figures are notoriously sensitive to input resolution, runtime, numeric precision, and target hardware, we deliberately separate the comparative analysis from measurement: the paper specifies a reproducible edge benchmarking protocol — fixed input regimes, warm/cold latency separation, sustained-throughput thermal testing, and energy-per-inference measurement — through which the framework's qualitative rankings can be instantiated as numbers on any given device. We conclude with deployment guidance matching model classes to application profiles and identify open problems in edge-aware neural architecture search and accuracy-preserving low-bit quantisation.

Keywords: Edge computing; Object detection; Lightweight deep learning; Model compression; Quantisation; Real-time inference; Embedded vision; MobileNet; YOLO; Benchmarking

DOI: 10.21590/ijhit.05.04.12

1. Introduction

Object detection has migrated from the data centre to the device. Traffic cameras count vehicles locally; drones track assets without a backhaul link; retail sensors detect shelf states in-store; industrial controllers inspect parts on the line. The motivations are consistent across domains — latency guarantees, bandwidth economics, privacy, and operation under intermittent connectivity — and they impose a common engineering regime: inference must complete within a frame budget, inside a memory envelope, at a power draw the enclosure can dissipate, on hardware costing tens rather than thousands of dollars.

Server-class detectors do not survive this regime. Two-stage architectures with heavy backbones deliver excellent accuracy at computational costs that embedded processors cannot meet at interactive rates. The response has been a decade of research into lightweight detection, proceeding along two complementary fronts: designing networks to be small (efficient backbones, single-stage heads, hardware-aware architecture search) and making trained networks smaller (quantisation, pruning, distillation). The resulting design space is crowded and confusing: model families overlap, reported benchmarks are measured under incompatible conditions, and practitioners routinely discover that a model's published desktop-GPU figures say little about its behaviour on their actual device. This paper addresses that confusion with three contributions: (i) a taxonomy of lightweight detection strategies (Section 2); (ii) a seven-dimension comparative framework applied across representative model families (Section 3, Table 1); and (iii) a reproducible benchmarking protocol that turns the framework into device-specific measurements, together with deployment guidance (Sections 4–5).

2. Taxonomy of Lightweight Detection Strategies

2.1 Efficient Backbone Design

The backbone — the convolutional feature extractor — dominates detector cost, and efficient backbone design has produced the field's foundational ideas. Depthwise separable convolutions factorise standard convolution into a per-channel spatial filter followed by a pointwise mixing layer, cutting multiply–accumulate operations by roughly an order of magnitude at modest accuracy cost; this is the signature of the MobileNet family. Inverted residual blocks with linear bottlenecks refine the idea by expanding, filtering, and re-compressing representations while preserving information flow through low-dimensional skip paths. Channel shuffling and grouped convolutions offer an alternative factorisation. Compound scaling treats depth, width, and input resolution as a jointly tuned budget rather than independent knobs, yielding families of models spanning a smooth accuracy–cost curve. Most recently, hardware-aware neural architecture search optimises directly for measured latency on a target processor rather than for proxy metrics such as parameter count — an important correction, since operations that are cheap in theory can be slow on real memory hierarchies.

2.2 Efficient Detection Heads

Single-stage detectors removed the region-proposal stage that made two-stage pipelines expensive, predicting classes and box offsets densely over feature maps in one pass; the SSD and YOLO lineages are the canonical examples. Multi-scale feature aggregation — feature pyramids and their lightweight bidirectional variants — recovers small-object accuracy that aggressive downsampling otherwise sacrifices, a critical property for edge use cases such as distant pedestrian or defect detection. Anchor-free head designs simplify the prediction machinery further, reducing both computation and the hyperparameter burden of anchor tuning.

2.3 Post-Design Compression

Orthogonal to architecture, compression shrinks trained models. Quantisation replaces 32-bit floating point with 8-bit integers (or lower), cutting memory and enabling the integer arithmetic units on which most edge accelerators achieve their throughput; quantisation-aware training recovers most of the accuracy that naive post-training conversion loses. Structured pruning removes whole channels or blocks, producing genuinely faster dense models, whereas unstructured pruning yields sparsity that commodity hardware often cannot exploit. Knowledge distillation trains the compact student to match a large teacher's output distribution, frequently recovering accuracy that architecture shrinkage alone forfeits. In mature deployments these techniques compose: a searched backbone, distilled during training, then quantised for the target runtime.

3. Comparative Framework and Analysis

Table 1 compares representative lightweight detection families across seven dimensions. Entries are deliberately qualitative — footprint and latency are expressed as relative classes rather than figures — because absolute numbers are functions of input resolution, numeric precision, runtime, and silicon, and transplanting published figures across those variables is precisely the malpractice this paper argues against. The framework's value is ordinal and structural: it tells a practitioner which families are candidates for a given constraint profile, to be resolved into numbers by the protocol of Section 4 on the actual target device.

Table 1. Qualitative comparison of representative lightweight object detection approaches

Model Family	Detection Paradigm	Efficiency Strategy	Relative Footprint	Latency Behaviour	Trade-off Profile	Hardware Affinity
MobileNet-SSD lineage	Single-stage, anchor-based	Depthwise separable / inverted residual backbone	Very small	Fast and stable at low resolutions	Gives up small-object accuracy first	Excellent on mobile CPUs and NPUs; mature INT8 paths
Compact YOLO lineage (tiny/nano variants)	Single-stage, dense prediction	Scaled-down backbone and head; aggregation retained	Small	Very fast; resolution-sensitive	Strong speed–accuracy balance for general scenes	Broad: GPUs, NPUs, DSPs; strong tooling ecosystem
Compound-scaled detectors (EfficientDet-style, small scales)	Single-stage with bidirectional feature pyramid	Compound scaling; efficient multi-scale fusion	Small–moderate	Moderate; deeper graphs cost latency	Best accuracy retention per compute at small scales	Favours accelerators with good graph compiler support
Hardware-searched detectors (NAS-derived)	Single-stage	Latency-in-the-loop architecture search	Very small–small	Optimised for the searched target	Excellent on-target; transfers poorly across silicon	Target-specific by construction
Quantised variants (INT8/INT4	Unchanged	Post-design compression	Reduced 2–4× vs. FP32	Large gains on integer accelerators	Small accuracy cost with quantisation-	Dependent on runtime and operator

Model Family	Detection Paradigm	Efficiency Strategy	Relative Footprint	Latency Behaviour	Trade-off Profile	Hardware Affinity
of any above)			parent		aware training	coverage
Distilled compact students	Unchanged	Teacher–student training	As student architecture	As student architecture	Recovers accuracy, not speed	Inherits student's affinity

3.1 Reading the Comparison

Three structural findings emerge. First, paradigm has converged: every viable edge detector is single-stage; the remaining differentiation lies in backbone strategy and multi-scale fusion. Second, the binding trade-off is small-object accuracy versus latency: the mechanisms that make models fast — low input resolution, aggressive downsampling, lean fusion — are exactly those that erase small and distant objects, so the application's object-size distribution should drive family selection more than headline accuracy metrics. Third, compression is now infrastructure, not exotica: integer quantisation is effectively mandatory to access modern edge accelerators, which makes runtime operator coverage and quantisation tooling maturity — the framework's final dimension — a first-class selection criterion that pure architecture comparisons ignore.

4. A Reproducible Edge Benchmarking Protocol

Published comparisons fail practitioners because they hold the wrong things constant. The following protocol specifies what a decision-grade, on-device evaluation must control. Fixed input regime: evaluate each candidate at the resolutions the application will actually use, not each model's flattering native resolution; report accuracy and latency per resolution. Precision parity: compare models in the numeric precision they will ship in, with quantisation-aware retraining applied uniformly. Latency decomposition: separate cold-start (model load, graph compilation) from warm steady-state latency, and report tail latencies, not means alone — frame-budget violations live in the tail. Sustained-throughput thermal testing: run continuous inference for an extended interval on the physically enclosed device and report the latency trajectory, since thermal throttling routinely erases short-burst benchmark advantages. Energy per inference: measure at the supply, since battery-bound deployments are governed by joules, not milliseconds. Accuracy on domain data: complement public benchmark evaluation with a held-out sample of application imagery, because domain shift affects compact models disproportionately. Full environment disclosure: device, firmware, runtime and compiler versions, and delegate/accelerator configuration, without which no reported number is reproducible. The protocol is model-agnostic and turns Table 1's ordinal guidance into defensible, device-specific measurements; representative results tables can then be populated per deployment rather than copied across incompatible contexts.

5. Deployment Guidance and Discussion

Mapping model classes to application profiles follows from the analysis. Ultra-constrained microcontroller-class targets push toward the smallest searched or MobileNet-derived

detectors with aggressive quantisation, accepting narrow object-class scopes. Battery-powered mobile and drone workloads favour compact YOLO or MobileNet-SSD variants at the lowest resolution the object-size distribution tolerates, with energy-per-inference as the ranking metric. Tethered accelerator-equipped nodes (industrial gateways, smart cameras with NPUs) can spend the additional budget on compound-scaled detectors or larger compact-YOLO scales to buy small-object accuracy, provided compiler support for the graph is verified early. Safety-adjacent uses (driver assistance, machinery guarding) should weight tail latency and thermal stability above mean throughput, and treat accuracy on domain data — not benchmark leaderboards — as the acceptance criterion. Two open problems merit research attention: architecture search that generalises across silicon families rather than overfitting one target, and low-bit (sub-8-bit) quantisation that preserves detection accuracy without per-device heroics. A limitation of the present study follows from its design: it supplies structure and protocol rather than new measurements, and its qualitative rankings should be validated — and will be sharpened — by protocol-conformant benchmarking campaigns across representative device classes.

6. Conclusion

Lightweight object detection has matured from a collection of clever architectures into an engineering discipline with converged design principles: single-stage heads, factorised or searched backbones, multi-scale fusion sized to the object distribution, and quantisation as standard infrastructure. This paper organised that discipline into a taxonomy and a seven-dimension comparative framework, argued that the decisive trade-offs are small-object accuracy versus latency and tooling maturity versus theoretical efficiency, and specified a benchmarking protocol that produces the device-specific numbers on which deployment decisions should actually rest. The practical message is a change of habit: choose the candidate set from structural analysis, but choose the model from measurements you made, on your device, under your thermal envelope, on your data. In edge vision, the benchmark that matters is the one that ships.

References

1. Desai, A. B., Rallabandi, B., Gattupalli, K. K., & Medarametla, M. S. (2025). Agentic AI for rural connectivity and spectrum-aware networks to power smart agriculture ecosystems. In 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICRTEECT67512.2025.11448879
2. Goyal, S., Rallabandi, B., Gattupalli, K. K., & Medarametla, M. S. (2025). Digital twin-enabled context-aware networks: Enhancing smart grid resilience through predictive intelligence. In 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICRTEECT67512.2025.11448880
3. Tripathi, N., Gattupalli, K. K., Rallabandi, B., & Medarametla, M. S. (2025). AI-native Cloud-RAN orchestration for enterprise private 5G using digital twin models. In 2025 1st IEEE Uttar Pradesh Section Women in Engineering International Conference on

Electrical Electronics and Computer Engineering (UPWIECON) (pp. 851–857). IEEE. https://doi.org/10.1109/UPWIECON67212.2025.11390348

4. Goyal, S., Gattupalli, K. K., Rallabandi, B., & Medarametla, M. S. (2025). Integrating terrestrial and non-terrestrial networks for reliable rural coverage in agriculture and smart grids. In 2025 1st IEEE Uttar Pradesh Section Women in Engineering International Conference on Electrical Electronics and Computer Engineering (UPWIECON) (pp. 846–850). IEEE. https://doi.org/10.1109/UPWIECON67212.2025.11390331
5. N. Pachauri, V. Suresh, M. V. V. Prasad Kantipudi, R. Alkanhel, and H. A. Abdallah, "Multi-Drug Scheduling for Chemotherapy Using Fractional Order Internal Model Controller," Mathematics, vol. 11, no. 8, Art. no. 1779, 2023,
6. M. V. V. Prasad Kantipudi, S. Vemuri, N. S. Pradeep Kumar, S. Sreenath Kashyap, and S. Eslamian, "Proposing Model for Water Quality Analysis Based on Hyperspectral Remote Sensor Data," in Handbook of Hydroinformatics, Elsevier, 2023, pp. 317–324.
7. Rana, M., Srinivas, S., Jamili, L. K., Jaiswal, I. A., Nakka, S., & Kasetti, S. (2025, May). Real-Time Monitoring and Prediction of Blood Sugar Levels in Diabetic Patients with Functional Models. In 2025 International Conference on Engineering, Technology & Management (ICETM) (pp. 1-6). IEEE.
8. Jaiswal, I. A. (2023). Intelligent Cybersecurity Framework for Large-Scale RESTful Service Architectures. International Journal of Research Radicals in Multidisciplinary Fields, ISSN: 2960-043X, 2 (1), 178–184. Retrieved from https://www. researchradicals. com/index. php/rr/article/view, 252.
9. Jaiswal, I. A. (2023). Multilingual and culturally adaptive AI models for global education platforms. International Journal for Research in Education (IJRE), 12(9), 17–27. https://doi.org/10.63345/ijre.v12.i9.1
10. M. S. Prashanth, R. Aluvalu, and M. V. V. Kantipudi, "Enhancing Health Product Traceability on the Blockchain: A Novel Approach for Supply Chain Management Inspection to AI," EAI Endorsed Transactions on Pervasive Health & Technology, vol. 10, no. 1, 2024.
11. H. K. Jani, M. V. V. Prasad Kantipudi, G. Nagababu, D. Prajapati, and S. S. Kachhwaha, "Simultaneity of Wind and Solar Energy: A Spatio-Temporal Analysis to Delineate the Plausible Regions to Harness," Sustainable Energy Technologies and Assessments, vol. 53, Art. no. 102665, 2022
12. Jaiswal, I. A. (2022). Natural language processing for security policy and log analysis. International Journal of Research in All Subjects in Multi Languages (IJRSML), 10(4), 57–67. https://doi.org/10.63345/ijrsml.v10.i4.1
13. Khan, S. (2019). Sales force security compliance: An in-depth study of GDPR, HIPAA, and PCI-DSS enforcement in cloud-based CRM systems and their implications for global enterprises. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 8(9), 2211–2219. https://doi.org/10.15662/IJAREEIE.2019.0809010
14. S. Choudhary, C. H. Kandikattu, S. Kumar, M. V. V. P. Kantipudi, and M. Kumar, "Enhancing cybersecurity through combined convolutional neural network-gated recurrent unit approach for distributed denial of service attack detection," in Proceedings of the 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), pp. 1–6, 2024.

15. S. Rani, D. Ghai, and S. Kumar, "Reconstruction of wire frame model of complex images using syntactic pattern recognition," in IET Conference Proceedings CP791, vol. 2021, no. 11, pp. 8–13, 2021.
16. Swathi, S. Kumar, S. Rani, A. Jain, and R. K. M. V. N. M., "Emotion classification using feature extraction of facial expression," in Proceedings of the 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 283–288, 2022.
17. S. Choudhary, S. Kumar, S. Gowroju, M. Gulhane, and R. Sri Lakshmi, Eds., Genomics at the Nexus of AI, Computer Vision, and Machine Learning. Hoboken, NJ, USA: John Wiley & Sons, 2024.
18. S. Kumar, S. Choudhary, S. Gowroju, and A. Bhola, "Convolutional neural network approach for multimodal biometric recognition system for banking sector on fusion of face and finger," in Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision, pp. 251–267, 2023.
19. S. Choudhary, S. Kumar, M. Gulhane, and M. Kumar, "Secured automated certificate creation based on multimodal biometric verification," in Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision, pp. 269–281, 2023.
20. Jaiswal, I. A. (2024). AI-powered observability and incident prediction in distributed enterprise platforms. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(1), 1–14. https://doi.org/10.63345/sjaibt.v1.i1.201
21. S. Choudhary, C. H. Kandikattu, S. Kumar, M. V. V. P. Kantipudi, and M. Kumar, "Enhancing cybersecurity through combined convolutional neural network-gated recurrent unit approach for distributed denial of service attack detection," in Proceedings of the 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), pp. 1–6, 2024.
22. Jaiswal, I. A. (2023). High-Performance AI-Augmented Content Management Systems for Distributed Clouds. International Journal of Multidisciplinary Innovation and Research Methodology, ISSN: 2960-2068, 2 (2), 90–97. Retrieved from https://ijmirm. com/index. php/ijmirm/article/view, 243.
23. Khan, S. (2018). The role of zero-trust models in database security: Eliminating implicit trust and enforcing continuous verification in enterprise data access systems. International Journal of Research in Electronics and Computer Engineering, 6(1), 1622–1628.
24. Jaiswal, I. A. (2024). Self-healing REST services using artificial intelligence in multi-cloud environments. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(3), 1–7. https://doi.org/10.63345/sjaibt.v1.i3.201
25. V. Saini, A. Jain, Anurag, and M. V. V. Prasad Kantipudi, "An Efficient Approach for Improving the Performance of Autonomous Vehicle Using Advanced Computer Vision," in International Conference on Soft Computing and Pattern Recognition, Cham, Switzerland: Springer Nature Switzerland, 2023, pp. 164–174.
26. Khan, S. (2017). The role of privacy-preserving techniques in database security: Secure multi-party computation, differential privacy, and homomorphic encryption. The Research Journal, 3(4), 29–35.
27. Khan, S. (2016). A study of data provenance and integrity in database security: Ensuring authenticity, non-repudiation, and accountability in data lifecycle management.

International Journal of Research in Electronics and Computer Engineering, 4(3), 180–186.

28. Cherladine, K., Rallabandi, B. R., Goel, A., Kaushik, K., Dhillon, A., & Soni, M. (2025). Generative adversarial defense models for simulated cyberattack scenarios in virtual networks. In 2025 International Conference on Digital Innovations for Sustainable Solutions (ICDISS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICDISS68238.2025.11320686
29. Rallabandi, B. R. (2020). MEC-native 5G systems: Orchestration algorithms for ultra-low latency cloud-edge integration. International Journal of Intelligent Systems and Applications in Engineering, 8(4), 398–408. https://ijisae.org/index.php/IJISAE/article/view/7889
30. Rallabandi, B. R. (2018). Joint deployment and operational energy optimization in heterogeneous cellular networks under traffic variability. International Journal of Communication Networks and Information Security, 10(3), 638–647. https://ijcnis.org/index.php/ijcnis/article/view/8604
31. C. H. Neetha and M. P. Kantipudi, "Adaptive Edge-Based Bilinear Interpolation for Smart Healthcare," in IET Conference Proceedings CP824, vol. 2022, no. 26, Stevenage, U.K.: The Institution of Engineering and Technology (IET), 2022, pp. 7–13.
32. R. Aluvalu, R. Venkatachalam, V. Uma Maheswari, R. J. Anandhi, M. V. V. Kantipudi, and S. Prashanth Mallellu, "IWHO-Based Cluster Head Selection for Vehicle-to-Vehicle Communication in Intelligent Transport System," International Journal of Transport Development and Integration, vol. 8, no. 2, pp. 154–166, 2024
33. S. Velamuri, S. K. Sudabattula, M. V. V. Prasad Kantipudi, and N. Prabaharan, "Q-Learning Based Commercial Electric Vehicles Scheduling in a Renewable Energy Dominant Distribution Systems," Electric Power Components and Systems, pp. 1–14, 2023
34. Singh, M., Rallabandi, B., Gattupalli, K. K., & Sai Medarametla, M. (2025). Autonomous private 5G field area networks: Achieving secure, scalable, and self-healing smart grid communications. In 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICRTEECT67512.2025.11448712.