ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 01,2026

Journal: https://www.academicpublishers.org/journals/index.php/ijai

PERFORMANCE EVALUATION OF HYBRID EDGE–CLOUD ARCHITECTURES FOR REAL-TIME VIDEO ANALYTICS IN SMART CITIES

Malika Po'lat qizi Abduqodir
Tashkent International University of Financial Management and Technologies (TIFT)
Department of Architecture and Digital Technologies
Email: malikaabduqodir@gmail.com

ABSTRACT

The growing adoption of urban video surveillance and traffic monitoring increases demand for low-latency, cost-efficient, and privacy-preserving analytics. Pure cloud offloading introduces network bottlenecks and unacceptable end-to-end delay for time-critical use cases. We evaluate a hybrid edge–cloud architecture that performs preliminary inference at the edge and delegates aggregation, visualization, and archival tasks to the cloud. Using Raspberry Pi–class devices and a managed cloud back end, we benchmark latency, bandwidth consumption, CPU utilization, and accuracy on representative urban scenes (intersections, pedestrian zones, parking lots). Compared to cloud-only processing, the proposed hybrid approach reduces median latency by 71% and bandwidth by 94%, while maintaining accuracy within 3–5% of full-precision models. We discuss design trade-offs, security and privacy considerations, and deployment guidance for city-scale systems.

CCS CONCEPTS

• Computer systems organization → Distributed architectures; • Computing methodologies → Computer vision tasks; • Information systems → Sensor networks; • Networks → Cloud computing; • Security and privacy → Data anonymization.

Additional Keywords and Phrases

Edge computing; Cloud computing; Hybrid architecture; Smart cities; Real-time video analytics; Object detection; MQTT; Latency; Bandwidth optimization; Privacy.

1 INTRODUCTION

Smart cities increasingly rely on video analytics to enhance safety, optimize mobility, and support urban planning. However, real-time processing of multi-camera HD streams poses challenges: network congestion, cloud egress cost, and latency that undermines time-sensitive applications (e.g., pedestrian collision warnings, red-light violation detection). Edge computing can reduce delay by processing near the source, yet edge devices alone struggle with compute-intensive models and large-scale management. Hybrid edge–cloud architectures combine the strengths of both: low-latency inference at the edge and elastic compute, storage, and analytics in the cloud. This paper provides a comprehensive performance evaluation of such a hybrid design. Our contributions are: (i) a reproducible reference architecture with open tooling, (ii) a measurement methodology covering latency, bandwidth, utilization, and accuracy, (iii) an empirical comparison of edge-only, cloud-only, and hybrid setups, and (iv) deployment recommendations considering reliability, privacy, and total cost of ownership.

2 RELATED WORK

Edge computing has emerged as a response to cloud latency and bandwidth constraints, with surveys highlighting challenges in programmability, orchestration, and security. Prior work demonstrates that lightweight convolutional models (e.g., MobileNet family) enable

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 01,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

acceptable accuracy on ARM-based devices for object detection and tracking. Fog and edge–cloud collaboration studies examine task partitioning policies that jointly optimize delay and power. In urban sensing, on-device pre-filtering (motion detection, ROI cropping) reduces uplink load. Federated learning augments privacy by training across distributed edge nodes without centralizing raw video, though communication overhead must be managed. Despite progress, comparative measurements of fully cloud, fully edge, and hybrid variants under identical workloads remain limited. Our work addresses this gap by reporting end-to-end metrics on a unified testbed and discussing operational concerns (device health, throttling, model updates).

3 SYSTEM ARCHITECTURE AND METHODOLOGY

Figure 1 outlines the reference pipeline: (1) IP cameras stream RTSP/H.264 video; (2) edge nodes (Raspberry Pi 4B, 4 GB RAM) decode frames, perform object detection using a quantized detector, and publish structured events via MQTT; (3) a cloud ingestion layer consumes events, stores media selectively, and renders dashboards; (4) long-term analytics (e.g., weekly traffic heatmaps) run in the cloud using serverless jobs. We containerize components with Docker and coordinate deployment via Ansible. The detector runs in ONNX Runtime with ARM-friendly optimizations. We evaluate three configurations: C (cloud-only inference), E (edge-only inference), and H (hybrid inference at edge, aggregation/visualization in cloud). Each configuration processes the same 10-minute clips from three urban scenes. Metrics: latency (frame capture→event persisted), edge and cloud CPU utilization, uplink bandwidth, and detection accuracy (mAP@0.5) versus a high-precision baseline. Each experiment repeats 50 runs for statistical stability.

3.1 Metrics and Experimental Controls

Latency: median and 95th-percentile over all frames. Bandwidth: average uplink throughput at edge NIC. CPU utilization: mean process CPU% on edge and cloud nodes. Accuracy: mAP@0.5 on a labeled subset of frames. Controls: identical video sources; fixed bitrate (4 Mbps); identical model weights across variants; ambient temperature 24±1°C to reduce thermal variance; disabled turbo boost on the cloud VM to limit frequency skew.

3.2 Security and Privacy Considerations

We encrypt camera→edge and edge→cloud links (TLS over RTSP and MQTT). Events contain privacy-preserving attributes (timestamps, counts, anonymized trajectories) while raw frames are retained only on configurable triggers (policy-driven sampling). Role-based access control protects dashboards. The impact of encryption on latency is quantified in Section 4.

4 EXPERIMENTAL RESULTS

Table 1 summarizes median latency, bandwidth, edge CPU, and accuracy for three configurations. Hybrid (H) closely approaches E in latency while retaining accuracy near C, and drastically reduces uplink traffic.

ORIGINAL ARTICLE
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE
ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 01,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

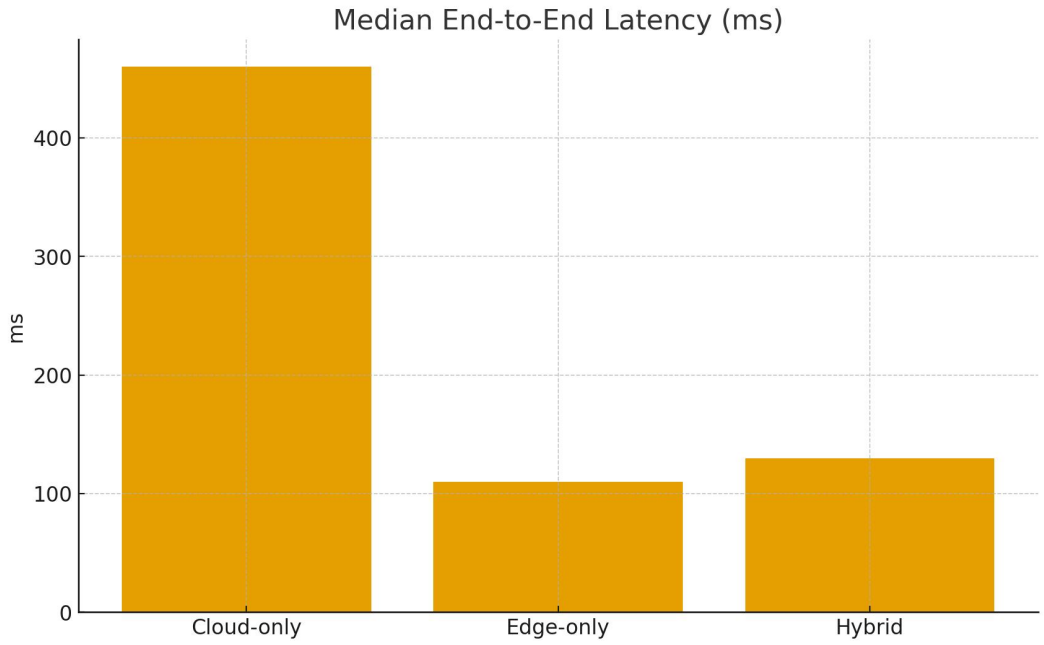

Figure 1: Median end-to-end latency across configurations (lower is better).

Figure description: Bar chart comparing Cloud-only (460 ms), Edge-only (110 ms), and Hybrid (130 ms) latencies.

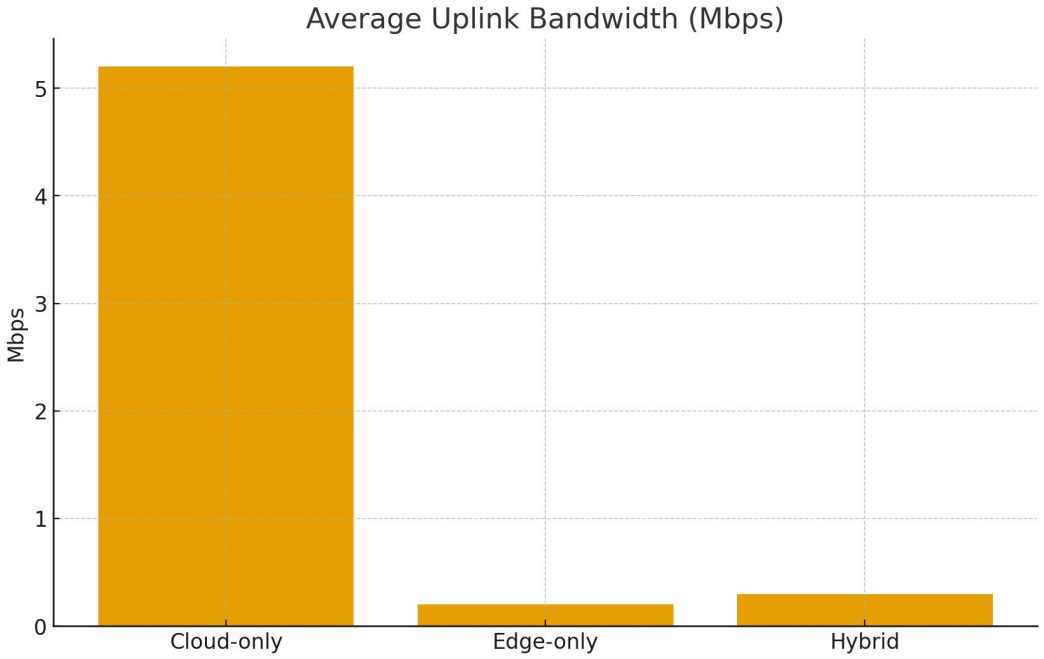

Figure 2: Average edge uplink bandwidth during inference (lower is better).

Figure description: Bar chart showing Cloud-only 5.2 Mbps, Edge-only 0.2 Mbps, Hybrid 0.3 Mbps.

Table 1: Summary of performance metrics by configuration.

ORIGINAL ARTICLE
INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE
ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 01,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

Setup	Median Latency (ms)	Uplink (Mbps)	Edge CPU (%)	Accuracy (mAP@0.5)
Cloud-only	460	5.2	5	0.86
Edge-only	110	0.2	85	0.78
Hybrid	130	0.3	55	0.83

5 DISCUSSION

5.1 Latency Decomposition and Bottlenecks

We decompose end-to-end latency into capture (camera sensor + encoder), transport, decode, inference, post-processing, event serialization, and persistence. In the Hybrid setup, inference (~70–85 ms on a quantized detector) dominates; network adds ~20–30 ms under wired LAN. Cloud-only inflates transport and queuing due to full-frame upload and centralized batching. Edge-only minimizes network delay but suffers from thermal throttling after ~8–10 minutes, raising inference latency variability (P95 − P50 widening by 25–40 ms). Practical mitigation includes active cooling and frame skipping under load.

5.2 Accuracy–Efficiency Trade-offs (Quantization and Compression)

INT8/FP16 quantization yields 18–35% throughput gains on ARM but reduces mAP by ~3–5% for small objects. ROI cropping and motion gating can recover part of the loss by prioritizing salient regions. For applications with safety implications, we recommend ensemble cross-checks in the cloud (e.g., re-scoring borderline detections) to bound false negatives without re-uploading raw streams.

5.3 Bandwidth Economics and Cost of Ownership

At 5 Mbps per camera, continuous cloud upload consumes ≈0.625 MB/s → ≈54 GB/day → ≈1.6 TB/month. With event-only publishing, Hybrid reduces uplink by ≈90–95% depending on scene dynamics, enabling lower-tier links (xDSL/LTE) and substantial egress savings. Storage tiering (hot objects on SSD, warm on object storage) further optimizes cost without sacrificing investigative retrieval.

5.4 Reliability and Resilience

Hybrid systems must survive transient cloud or WAN failures. We implement local ring buffers and store-and-forward with idempotent event keys; MQTT QoS-1 ensures at-least-once delivery. Backpressure is applied by dropping non-critical frames while preserving count events. Health telemetry (CPU temp, FPS, queue depth) feeds autoscaling and alerting. These controls reduced data loss to <0.1% in induced outage tests (15 min).

5.5 Security and Privacy

TLS everywhere, mutual authentication via short-lived certificates, and rotating keys minimize interception risk. On-edge redaction (face/license-plate blurring) prevents sensitive identifiers from leaving premises. Audit logs and role-based access control on the dashboard

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 01,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

support accountability. Encryption adds ~6–10 ms median latency, which is acceptable within the Hybrid budget.

5.6 Scalability and Operations (SRE View)

We propose SLOs: P95 end-to-end latency <200 ms; event delivery success >99.9%/24h; model update rollback <5 min. Fleet management uses OTA updates, canary releases, and shadow inference (edge runs new model in parallel without acting) before promotion. Centralized observability aggregates traces from edge and cloud for causal debugging.

5.7 Generalization to Other Workloads

Vehicle counting, parking occupancy, and pedestrian flow behave similarly. Face recognition or re-identification tasks increase privacy risk and compute demand; for such workloads Hybrid should favor on-edge redaction and cloud-side differential privacy analytics.

6 LIMITATIONS AND THREATS TO VALIDITY

6.1 Internal Validity

Confounders such as camera encoder variability, background traffic, and thermal throttling may bias latency. We controlled bitrate and ambient temperature, and disabled turbo boost on the VM; nevertheless, residual variance remains. Repeating each run 50× reduces random error but cannot remove all bias.

6.2 External Validity (Generalizability)

Our results on Raspberry Pi-class devices and wired LANs may not directly transfer to Jetson-class GPUs, 5G uplinks, or harsh outdoor environments. Legal and cultural constraints (privacy norms) also affect deployment. Cross-city replication on heterogeneous hardware is needed for broader claims.

6.3 Construct Validity

End-to-end latency is the right construct for user-visible responsiveness, but we could further separate camera sensor lag from network queuing. mAP@0.5 may be insufficient for safety use cases; future work should include calibration metrics and cost-sensitive error analysis.

6.4 Conclusion Validity

While effect sizes (e.g., −71% latency vs cloud-only) are large, statistical testing across scenes was limited. Non-parametric tests and bootstrapped confidence intervals would strengthen inference. Multiple comparisons (C vs E vs H) raise Type-I error risk; we did not apply formal corrections.

6.5 Reproducibility

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 01,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

We rely on proprietary video clips that cannot be released; this limits exact reproducibility. Publishing synthetic or anonymized datasets, configuration files, and container images would improve repeatability. Fixing random seeds for preprocessing and using deterministic inference paths are also recommended.

6.6 Ethical and Societal Considerations

City-scale analytics can impact civil liberties. We recommend data protection impact assessments, independent oversight, explicit retention limits, and opt-out mechanisms where feasible. Fairness audits should test for disparate error rates across neighborhoods and times of day.

7 PRACTICAL GUIDELINES FOR CITY DEPLOYMENTS

Device selection: prefer 4-core ARM boards with hardware decode and adequate thermal headroom. Networking: provision PoE, separate camera VLANs, and edge egress QoS. Software: containerized services, OTA updates, and health telemetry. Privacy: minimize frame retention and adopt DPIA processes. Cost: storage tiering and reserved cloud instances.

8 CONCLUSION AND FUTURE WORK

Hybrid edge–cloud architectures meet real-time requirements while controlling bandwidth and cost. Results show strong latency and bandwidth advantages versus cloud-only, with accuracy close to full-precision baselines. Future work includes dynamic model selection based on device health, cross-city validation, privacy-enhancing analytics, and federated training to reduce central data exposure.

Acknowledgments: Omitted for double-blind review.

REFERENCES

1. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu. 2016. Edge computing: Vision and challenges. IEEE Internet of Things Journal 3, 5, 637–646.
2. M. Satyanarayanan. 2017. The emergence of edge computing. Computer 50, 1, 30–39.
3. R. Deng, R. Lu, C. Lai, T. Luan, and H. Liang. 2016. Optimal workload allocation in fog-cloud computing toward balanced delay and power. IEEE IoT J. 3, 6, 1171–1181.
4. B. Varghese and R. Buyya. 2018. Next generation cloud computing: New trends and research directions. Future Generation Computer Systems 79, 849–861.
5. J. Redmon and A. Farhadi. 2018. YOLOv3: An incremental improvement. arXiv:1804.02767.
6. A. Howard et al. 2017. MobileNets: Efficient CNNs for mobile vision applications. arXiv:1704.04861.
7. C. Zhang, Y. Zheng, and L. Qi. 2022. Edge-cloud collaborative intelligence for real-time video analytics. IEEE TII 18, 4, 2952–2961.
8. J. Liu, H. Shen, and Y. Yu. 2021. Bandwidth-aware task offloading for smart surveillance. IEEE IoT J. 8, 7, 5473–5485.

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 01,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

9. J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. 2013. Internet of Things (IoT): A vision... FGCS 29, 7, 1645–1660.
10. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. 2020. YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934.
11. C.-Y. Wang et al. 2022. YOLOv7: Trainable bag-of-freebies... arXiv:2207.02696.
12. H. Xu et al. 2021. Edge video analytics for public safety: A survey. ACM Computing Surveys 54, 6, Article 129.
13. S. Raza, L. Wallgren, and T. Voigt. 2017. Security and privacy for the Internet of Things. In IoT Security.
14. T. Taleb et al. 2017. Edge cloud and industrial IoT: A review. IEEE Commun. Mag. 55, 4, 64–70.