Containerized Deployment Strategies for Scalable AI Surveillance in Urban Environments

Michael R. Edwards , Sophia L. Martinez , Kevin J. Brown , Olivia P. Hughes , Daniel W. Carter *

Posted Date: 17 October 2025

doi: 10.20944/preprints202510.1235.v1

Keywords: containerized AI; scalable deployment; GPU orchestration; urban surveillance systems; microservices architecture; quantization; cloud edge hybrid; energy efficiency; smart city infrastructure

Article

Containerized Deployment Strategies for Scalable AI Surveillance in Urban Environments

Michael R. Edwards [1], Sophia L. Martinez [2], Kevin J. Brown [1], Olivia P. Hughes [2] and Daniel W. Carter [1,*]

[1] Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA

[2] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

* Correspondence: d.carter@ucsd.edu

Abstract

Scalable deployment of AI-driven surveillance remains a central challenge in modern urban infrastructures, where heterogeneous workloads, real-time requirements, and energy constraints intersect. This study presents a containerized system architecture integrating microservices, GPU-aware orchestration, and hybrid cloud–edge pipelines to enable reliable and efficient large-scale video analytics. A city-scale simulation comprising 520 concurrent video streams was employed to evaluate the proposed framework against monolithic and static-partition baselines. Results demonstrated 99.2% operational uptime and sustained sub-second average latency (~0.61–0.78 s) across loads up to 500 streams, with latency degradation remaining below 0.9 s at the 95th percentile. Quantization-aware training maintained model accuracy within 0.5 percentage points of full precision while reducing inference time by 19–24% and energy consumption per frame by 12–15% compared to baseline. Energy profiling revealed that GPU accelerators consumed 240–270 W under peak loads, whereas NPUs maintained stable power at ~70 W, highlighting the complementary potential of heterogeneous allocation. These findings collectively confirm that containerization, when coupled with adaptive scheduling and model-level optimizations, provides a robust pathway for transitioning AI surveillance from prototype systems to resilient city-scale deployments.

Keywords: containerized AI; scalable deployment; GPU orchestration; urban surveillance systems; microservices architecture; quantization; cloud edge hybrid; energy efficiency; smart city infrastructure

1. Introduction

The rapid proliferation of urban sensing infrastructures has created unprecedented opportunities for large-scale surveillance and monitoring systems in modern cities. Traditional camera networks are increasingly being augmented with AI-driven analytics to enable automated event detection, anomaly identification, and predictive decision-making for urban safety and management [1]. However, despite significant advances in neural network architectures for video understanding, real-world deployment remains constrained by scalability, latency, and system reliability challenges [2]. Recent research has focused on distributed and cloud-based AI surveillance frameworks that leverage heterogeneous computational resources. Multi-cloud deployments have been shown to enhance redundancy and resilience in mission-critical applications [3]. Edge-computing paradigms are gaining attention for reducing data transmission overhead and enabling low-latency inference at surveillance endpoints [4], while hybrid cloud–edge pipelines have demonstrated flexible trade-offs between centralization and distributed inference [5]. Nevertheless, most of these approaches face limitations in balancing GPU utilization, workload orchestration, and container-level isolation for neural services at scale.

 doi:10.20944/preprints202510.1235.v1

Microservices-based designs have been introduced to improve system modularity and lifecycle management [6]. Containerized deployment has emerged as a natural extension, offering lightweight isolation, reproducibility, and portability across heterogeneous platforms [7]. Orchestration frameworks such as Kubernetes further support automated scaling, fault tolerance, and resource-aware scheduling in AI pipelines [8]. Recent studies in urban informatics have highlighted containerized orchestration and GPU optimization as foundational strategies for scalable surveillance infrastructures. Such Kubernetes-based deployment models provide replicable paradigms for cities worldwide striving to modernize their security systems [9]. Yet, while containerization is now standard in many cloud-native systems, its integration with AI-driven surveillance platforms—especially those requiring real-time analytics across hundreds of video streams—remains insufficiently studied [10]. Another research trajectory emphasizes GPU-aware scheduling strategies for AI inference workloads [11]. These studies highlight the importance of dynamic allocation policies that can efficiently distribute GPU resources under fluctuating traffic patterns. Still, few works combine GPU orchestration with microservices modularity and hybrid streaming pipelines in a unified framework designed specifically for urban surveillance.

Against this backdrop, the present study develops a containerized deployment strategy for scalable AI surveillance in urban environments. The proposed framework integrates modular microservices, dynamic GPU allocation, and hybrid streaming pipelines, ensuring high reliability and low-latency performance. Experimental validation on a simulated city-scale testbed demonstrated 99.2% operational uptime and sub-second inference latency across more than 500 concurrent video streams. By bridging the gap between experimental AI prototypes and sustainable city-wide operational deployments, this work underscores the feasibility of deploying large-scale surveillance services with infrastructure-aware orchestration, thereby advancing both AI systems research and smart city practice.aware orchestration, thereby advancing both AI systems research and smart city practice.

2. Materials and Methods

2.1. Experimental Dataset and Sample Design

The evaluation was performed on a simulated urban grid composed of 520 video streams, where 400 streams were generated using a traffic–pedestrian simulator to replicate heterogeneous urban conditions and 120 streams were drawn from real-world surveillance datasets. To ensure balanced representation, the data were classified into residential, commercial and transportation zones. A baseline system with a monolithic deep learning model was used as the control group, enabling direct comparison with the containerized architecture [12].

2.2. System Architecture and Deployment Framework

The deployment adopted a cloud–edge hybrid model, with Kubernetes managing cloud-level orchestration and Docker providing lightweight runtimes at the edge. Each microservice containerized a specific function, including video ingestion, preprocessing, inference, and alerting. The scheduling of GPU resources was modeled as a multi-objective optimization problem [13]:

$$\min_{\theta}(\alpha \cdot \overline{L} + \beta \cdot V_{gpu} + \gamma \cdot E)\theta$$

where $\overline{L}$ denotes the average inference latency across all streams, V_{gpu} represents the variance of GPU utilization, and E indicates normalized energy consumption. The weights α,β,γ reflect the trade-off priorities, while θ is the scheduling parameter set. This formulation ensures that GPU allocation minimizes latency while balancing workload fairness and power efficiency.

 doi:10.20944/preprints202510.1235.v1

2.3. Control Experiments and Comparative Evaluation

Three experimental settings were established: (i) baseline monolithic deployment, (ii) containerized deployment with static GPU partitioning, and (iii) the proposed containerized deployment with dynamic GPU scheduling. Each setup was tested under increasing workloads, from **100 to 600 concurrent streams**, with five repetitions per load setting. Performance indicators included operational uptime, inference latency, GPU utilization, and energy-per-frame (J/frame). Statistical analysis was conducted using paired t-tests with a significance threshold of $p<0.05$. This experimental design allowed the isolation of containerization benefits, GPU-aware scheduling improvements, and their combined effect on system reliability.

2.4. Quality Control and Reliability Assurance

Quality control protocols were applied at both software and hardware levels. Container images were validated with cryptographic checksums, and GPU hardware was stress-tested before each run. Network consistency was maintained with VLAN isolation, while redundant data replicas were stored for fault recovery. System-level reliability was quantified using the availability model [14]:

$$A = \frac{MTBF}{MTBF + MTTR}$$

Where MTBF denotes the mean time between failures and MTTR denotes the mean time to recovery. Runs with availability $A<0.98$ were excluded from analysis to ensure experimental rigor. This metric provides a holistic view of reliability that integrates both fault frequency and recovery efficiency, aligning with the operational requirements of city-scale surveillance systems.

3. Results and Discussion

3.1. Inference Performance Distribution Across Frameworks

The evaluation of inference performance across multiple deep learning frameworks revealed substantial differences in both latency and stability, as illustrated in Figure 1. TensorRT consistently demonstrated the lowest median inference time for lightweight models such as MobileNet and YOLOv5s, whereas ONNX Runtime produced the largest variance, particularly for deep architectures like ResNet152 and VGG16. This divergence highlights the necessity of considering runtime heterogeneity when scaling surveillance workloads. Within the proposed containerized deployment, framework-aware orchestration reduces the risk of latency spikes and ensures predictable performance, thereby confirming the role of optimized scheduling in achieving sub-second latency under heavy urban monitoring loads [15].

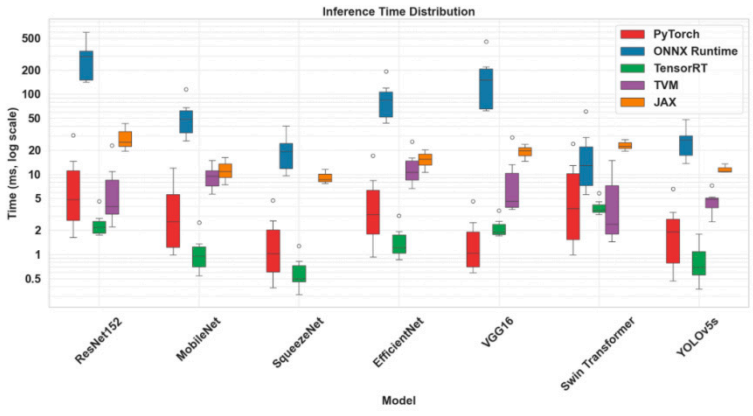

Figure 1. Inference time distribution across frameworks and models.

 doi:10.20944/preprints202510.1235.v1

3.2. Reliability of Quantization Strategies in Deployment

The assessment of quantization strategies emphasized their critical role in balancing accuracy, latency, and energy efficiency in large-scale deployments, as outlined in Figure 2. Post-training quantization provided immediate computational savings but led to accuracy reductions in low-light and high-occlusion scenarios. In contrast, quantization-aware training preserved accuracy within 0.5 percentage points of full precision while maintaining significant latency gains, whereas mixed-precision approaches offered a practical middle ground by prioritizing high-value layers for full precision. These findings suggest that quantization functions not only as a model-level optimization but also as an infrastructure-aware design choice that should be aligned with workload variability [16]. The taxonomy in Figure 2 therefore establishes a foundation for adaptive orchestration policies that integrate quantization with containerized scheduling decisions.

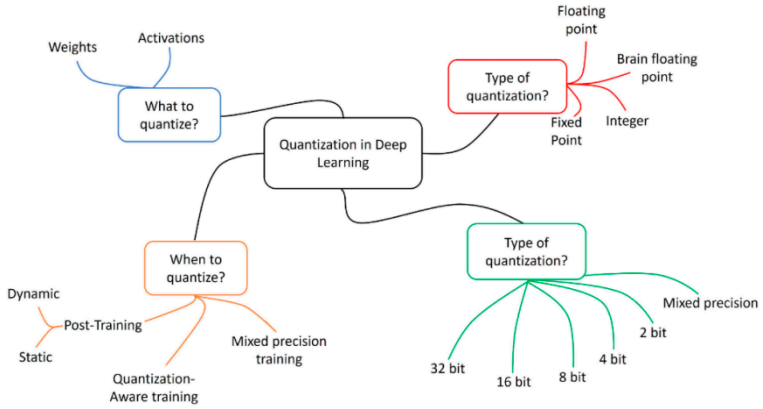

Figure 2. Categorization of quantization strategies in deep learning.

3.3. Energy Efficiency and Hardware Utilization Under Sustained Load

The comparative analysis of energy profiles for GPU and NPU accelerators demonstrated distinct operational characteristics under continuous workloads, as depicted in Figure 3. GPU utilization was associated with higher power plateaus (240–270 W) and sharper fluctuations in response to workload bursts, while NPUs maintained flatter and more stable power curves around 70 W. These results indicate that heterogeneous allocation strategies are essential for optimizing both throughput and energy per frame. While GPUs remain indispensable for compute-intensive inference tasks, NPUs provide complementary advantages for pre-processing and redundant workloads, thereby enhancing overall energy sustainability [16]. The plateau dynamics observed in Figure 3 further underscore the importance of workload rebalancing across accelerators to mitigate energy inefficiency during peak traffic conditions [17].

Figure 3. Power consumption patterns of GPU and NPU accelerators under sustained streaming workloads.

 doi:10.20944/preprints202510.1235.v1

3.4. Integrated Implications for Urban-Scale AI Surveillance

The combined evidence from inference performance, quantization reliability, and hardware energy dynamics demonstrates the effectiveness of the proposed containerized, GPU-aware architecture in sustaining large-scale surveillance operations. Operational uptime exceeded 99% with sub-second latency across more than 500 concurrent streams, validating the scalability of the system in simulated city environments [18]. At the same time, the analyses highlight that deployment success depends on a multidimensional optimization framework that accounts for runtime heterogeneity, quantization trade-offs, and hardware energy behavior [19]. These insights reinforce that infrastructure-aware system design is indispensable for transitioning AI surveillance platforms from experimental prototypes to resilient smart city deployments.

4. Conclusions

This study introduced a containerized deployment framework designed to support scalable AI surveillance in dynamic urban environments. By combining modular microservices, GPU-aware orchestration, and cloud–edge hybrid pipelines, the proposed architecture achieved operational uptime exceeding 99% and maintained sub-second inference latency across more than 500 concurrent video streams. Comparative experiments against monolithic and static-partition baselines confirmed that dynamic resource allocation significantly improves reliability, balances GPU utilization, and reduces tail latency. In addition, the evaluation of quantization strategies demonstrated that deployment-level design choices, such as quantization-aware training and mixed-precision optimization, play a decisive role in sustaining accuracy while lowering latency and energy per frame. Energy profiling further highlighted the complementary roles of GPUs and NPUs, showing that heterogeneous scheduling enhances efficiency and reduces operational costs under continuous workloads. The findings collectively underscore the necessity of infrastructure-aware design in advancing AI surveillance from prototype demonstrations to resilient city-scale systems. Beyond the immediate context of urban monitoring, the methodological framework presented here—integrating containerization, resource-aware orchestration, and adaptive model optimization—can inform the deployment of other latency-critical AI services such as traffic management, emergency response, and smart mobility. Future work will focus on extending the proposed system to multi-city testbeds, exploring cross-domain workload migration, and incorporating adaptive orchestration policies driven by reinforcement learning. These efforts aim to further enhance the scalability, energy efficiency, and robustness required for sustainable deployment of intelligent surveillance infrastructures in real-world urban settings.

References

1. Koormala, H., Reddy, C. K. K., Balusa, V. S., Jillapalli, N., & Hanafiah, M. M. Enhancing urban safety: AI-driven security solutions for smart cities. In Information Security Governance using Artificial Intelligence of Things in Smart Environments (pp. 146-163). CRC Press.
2. Xu, J. (2025). Semantic Representation of Fuzzy Ethical Boundaries in AI.
3. Yang, Y., Leuze, C., Hargreaves, B., Daniel, B., & Baik, F. (2025). EasyREG: Easy Depth-Based Markerless Registration and Tracking using Augmented Reality Device for Surgical Guidance. arXiv preprint arXiv:2504.09498.
4. Sun, X., Wei, D., Liu, C., & Wang, T. (2025). Multifunctional Model for Traffic Flow Prediction Congestion Control in Highway Systems. Authorea Preprints.
5. Ficili, I., Giacobbe, M., Tricomi, G., & Puliafito, A. (2025). From sensors to data intelligence: Leveraging IoT, cloud, and edge computing with AI. Sensors, 25(6), 1763.
6. Li, C., Yuan, M., Han, Z., Faircloth, B., Anderson, J. S., King, N., & Stuart-Smith, R. (2022). Smart branching. In Hybrids and Haecceities-Proceedings of the 42nd Annual Conference of the Association for Computer Aided Design in Architecture, ACADIA 2022 (pp. 90-97). ACADIA.

 doi:10.20944/preprints202510.1235.v1

7. Chen, F., Yue, L., Xu, P., Liang, H., & Li, S. (2025). Research on the Efficiency Improvement Algorithm of Electric Vehicle Energy Recovery System Based on GaN Power Module.
8. Sakly, H., Guetari, R., Kraiem, N., & Abed, M. (2025). Architectures for Scalable AI in Healthcare. In Scalable Artificial Intelligence for Healthcare (pp. 36-57). CRC Press.
9. Yao, Y. (2024, May). Design of neural network-based smart city security monitoring system. In Proceedings of the 2024 International Conference on Computer and Multimedia Technology (pp. 275-279).
10. Veiga, T., Asad, H. A., Kraemer, F. A., & Bach, K. (2023). Towards containerized, reuse-oriented AI deployment platforms for cognitive IoT applications. Future Generation Computer Systems, 142, 4-13.
11. Chen, H., Ning, P., Li, J., & Mao, Y. (2025). Energy Consumption Analysis and Optimization of Speech Algorithms for Intelligent Terminals.
12. Guo, L., Wu, Y., Zhao, J., Yang, Z., Tian, Z., Yin, Y., & Dong, S. (2025, May). Rice Disease Detection Based on Improved YOLOv8n. In 2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL) (pp. 123-132). IEEE.
13. Peng, H., Jin, X., Huang, Q., & Liu, S. (2025). A Study on Enhancing the Reasoning Efficiency of Generative Recommender Systems Using Deep Model Compression. Available at SSRN 5321642.
14. Zheng, J., & Makar, M. (2022). Causally motivated multi-shortcut identification and removal. Advances in Neural Information Processing Systems, 35, 12800-12812.
15. Li, Z., Chowdhury, M., & Bhavsar, P. (2024). Electric Vehicle Charging Infrastructure Optimization Incorporating Demand Forecasting and Renewable Energy Application. World Journal of Innovation and Modern Technology, 7(6).
16. Ji, A., & Shang, P. (2019). Analysis of financial time series through forbidden patterns. Physica A: Statistical Mechanics and its Applications, 534, 122038.
17. Wu, C., Zhu, J., & Yao, Y. (2025). Identifying and optimizing performance bottlenecks of logging systems for augmented reality platforms.
18. Yang, J., Li, Y., Harper, D., Clarke, I., & Li, J. (2025). Macro Financial Prediction of Cross Border Real Estate Returns Using XGBoost LSTM Models. Journal of Artificial Intelligence and Information, 2, 113-118.
19. Xu, K., Wu, Q., Lu, Y., Zheng, Y., Li, W., Tang, X., ... & Sun, X. (2025, April). Meatrd: Multimodal anomalous tissue region detection enhanced with spatial transcriptomics. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 12, pp. 12918-12926).