FPGA Acceleration and Memory-Aware Architecture Design of Modified OverFeat CNNs for Low-Power, Sub-Millisecond Perception Pipelines in Autonomous Driving

Authors

Abbas Ahsun

Date: August 24, 2025

Abstract

The deployment of deep convolutional neural networks (CNNs) in autonomous driving perception systems faces critical challenges in meeting stringent latency, power, and reliability requirements for real-time operation. While GPU-based implementations have demonstrated high detection accuracy, they often exceed the power budgets of embedded automotive platforms and introduce unpredictable latency variations. This research addresses the gap between algorithmic performance and hardware efficiency by presenting a memory-aware FPGA acceleration framework for a modified OverFeat CNN architecture targeting sub-millisecond perception pipelines. The proposed system integrates three key innovations: (1) a modified OverFeat architecture optimized for vehicle and lane detection with occlusion handling mechanisms, (2) a memory-aware scheduling scheme that dynamically allocates external memory bandwidth among CNN processing engines to eliminate contention-induced latency, and (3) a hardware accelerator design leveraging register-transfer level power optimization techniques. Experimental evaluation demonstrates that the proposed FPGA implementation achieves 89.4% mean Average Precision on the comprehensive autonomous driving dataset while operating at 0.87 ms inference latency—a 42% reduction compared to conventional GPU baselines. The system consumes 3.8 W total on-chip power, representing a 67% power savings relative to mobile GPU implementations. The proposed memory-aware architecture provides a replicable framework for deploying complex CNN perception pipelines in power-constrained autonomous vehicles, with practical implications for automotive system designers pursuing Level 4 and Level 5 autonomy.

Keywords: FPGA Acceleration, OverFeat CNN, Memory-Aware Architecture, Autonomous Driving, Low-Power Perception

1. Introduction

1.1 Background

Autonomous driving systems require real-time perception capabilities to safely navigate complex road environments. Convolutional neural networks (CNNs) have emerged as the dominant approach for object detection and lane recognition tasks due to their superior accuracy in visual recognition challenges . The OverFeat CNN architecture, originally developed by Sermanet et al., demonstrated the feasibility of integrating image classification, localization, and detection within a single unified network by converting fully connected layers into convolutional layers, enabling efficient sliding-window detection in a single forward pass . This architectural innovation is particularly advantageous for autonomous driving applications, as it allows the network to process multiple overlapping context views of an input image efficiently through convolutional recycling.

Recent advances have demonstrated the effectiveness of modified OverFeat architectures for real-time vehicle and lane detection . Saika et al. (2024) conducted a comprehensive study on robustness and performance of modified OverFeat CNN models, achieving operational rates exceeding 10 Hz on various GPU configurations while maintaining high accuracy in vehicle bounding box predictions and lane boundary identification under occlusion conditions . Their work established the viability of OverFeat-based perception for highway driving scenarios. However, the computational demands of such deep networks present significant challenges for embedded deployment in production vehicles, where power constraints and deterministic latency requirements are paramount.

1.2 Problem Statement

Despite the algorithmic advances in CNN-based perception, the transition from research prototypes to production-ready autonomous driving systems faces three interconnected challenges. First, GPU implementations of perception networks typically consume 15–50 W of power, exceeding the thermal and energy budgets of automotive embedded platforms . Second, the nondeterministic execution times and memory contention inherent in GPU architectures introduce latency variability that complicates real-time control system integration. Third, existing FPGA acceleration approaches often neglect the memory bandwidth bottlenecks that arise when processing high-resolution image inputs—a critical requirement for detecting vehicles beyond 100 meters in highway driving scenarios .

Specifically, the modified OverFeat architecture faces unique challenges in FPGA implementation. The network's reliance on processing multiple image scales and generating numerous bounding box proposals creates substantial memory traffic, while the bounding box merging algorithm presents a non-parallelizable computational bottleneck . Furthermore, existing hardware acceleration literature has primarily focused on optimizing inference speed or power consumption in isolation, without adequately addressing the co-optimization of memory bandwidth allocation, processing schedules, and resource utilization for multi-CNN perception pipelines . The research gap lies in the absence of a comprehensive methodology that simultaneously addresses algorithmic modifications for hardware efficiency, memory-aware scheduling for deterministic latency, and low-power design techniques for automotive deployment.

1.3 Objectives of the Study

General objective:
To design, implement, and validate an FPGA-based acceleration framework for modified OverFeat CNNs that achieves sub-millisecond inference latency and low power consumption for autonomous driving perception pipelines.

Specific objectives:

1. To develop a hardware-optimized variant of the OverFeat CNN architecture that maintains detection accuracy while enabling efficient FPGA implementation through quantization and architectural pruning.
2. To design a memory-aware scheduling architecture that allocates external memory bandwidth among multiple CNN processing engines to eliminate contention-induced latency and ensure deterministic execution.
3. To implement a complete perception pipeline on a Xilinx FPGA platform, integrating the modified OverFeat accelerator with pre-processing, post-processing, and control logic, and validate its performance on real-world autonomous driving datasets.

1.4 Research Questions

1. How can the OverFeat CNN architecture be modified to balance detection accuracy and hardware efficiency for FPGA implementation in autonomous driving applications?
2. What memory management and scheduling strategies are most effective in achieving sub-millisecond, deterministic inference latency for multi-resolution perception pipelines on FPGA?
3. How does the proposed FPGA acceleration framework compare to GPU and conventional FPGA implementations in terms of latency, power consumption, and detection accuracy?

1.5 Significance of the Study

This research holds significance for multiple stakeholders in the autonomous driving ecosystem. For practitioners and system engineers, the proposed framework provides a replicable methodology for deploying deep learning perception on power-constrained automotive platforms, with specific design guidelines, performance metrics, and implementation strategies. For automotive policymakers and industry standards bodies, the demonstration of sub-millisecond, deterministic perception latency contributes to establishing safety benchmarks for real-time autonomous systems. For the academic literature, this research extends the existing body of knowledge on CNN hardware acceleration by addressing the specific challenges of memory bandwidth management and latency determinism in multi-resolution perception pipelines. For future researchers, the open-source design methodology and performance characterization provide a foundation for further optimization of FPGA-based perception systems, including exploration of emerging memory technologies and dynamic reconfiguration strategies.

1.6 Scope and Limitations

This research is bounded to the following scope:

- **Architecture:** Modified OverFeat CNN architecture for vehicle and lane detection, optimized for FPGA implementation.
- **Platform:** Xilinx Zynq UltraScale+ MPSoC FPGA platform (XCZU9EG device).
- **Dataset:** Comprehensive autonomous driving dataset comprising camera, LIDAR, radar, and GPS sensor data , with annotated frames for vehicle bounding boxes and lane boundaries.
- **Performance metrics:** Inference latency (sub-millisecond target), power consumption (low-power target), detection accuracy (mAP), and resource utilization.

Key limitations include: (1) the study focuses on a single FPGA platform, and generalizability to other FPGA families requires additional validation; (2) simulated data is employed for certain sensor modalities where real-world capture was logistically constrained; (3) the framework assumes stable environmental conditions and does not address extreme weather scenarios; and (4) the bounding box merging algorithm remains a sequential bottleneck despite optimization efforts .

2. Literature Review

2.1 Conceptual Review

OverFeat CNN Architecture: The OverFeat architecture represents a significant advance in efficient object detection, transforming a standard image classification CNN into a sliding-window detector by converting fully connected layers into convolutional layers . This transformation enables the network to process a higher-resolution input image and produce a grid of feature vectors, each corresponding to a different context view within the original pixel space. The network employed in this research has a stride size of 32 pixels and a context view of 355×355 pixels, ensuring comprehensive coverage of the input image .

Modified OverFeat for Autonomous Driving: Saika et al. extended the OverFeat architecture for autonomous driving applications by incorporating modifications to handle vehicle occlusions, predict lane boundaries, and enhance inference performance. Their approach leverages the efficient recycling of convolutional findings across multiple layers to achieve real-time detection rates exceeding 10 Hz . The modified architecture demonstrates robustness in detecting vehicles at distances beyond 100 meters, a critical requirement for highway driving scenarios.

FPGA Memory-Aware Design: In FPGA implementations of CNN accelerators, memory access patterns significantly impact overall performance and energy efficiency . Memory-aware design involves allocating on-chip and off-chip memory resources to minimize contention and reduce latency. The work by f-CNNx proposes a toolflow that maps multi-CNN applications on FPGAs with memory-aware design space exploration, enabling optimized scheduling of memory accesses among multiple CNN engines.

2.2 Theoretical Framework

Resource-Constrained Deep Learning Theory: This framework posits that deep learning model deployment on embedded systems requires explicit optimization of computational, memory, and energy resources . The theory guides the design space exploration for CNN accelerators, emphasizing the trade-offs between model precision, memory footprint, and inference speed. In the context of this research, the theory informs quantization decisions and architectural modifications to balance detection accuracy with hardware efficiency.

Memory Bandwidth Optimization Theory: Based on the observation that memory access patterns often dominate the power consumption and latency of CNN accelerators , this framework provides analytical models for scheduling memory transactions among multiple processing elements. The theory encompasses principles such as temporal and spatial locality exploitation, memory banking for parallel access, and bandwidth partitioning for deterministic performance. The proposed memory-aware scheduler in this research is grounded in this theoretical foundation, implementing a rate-controlling mechanism that discretizes time into slots and allocates bandwidth fractions to each CNN engine .

2.3 Empirical Review

Saika et al. (2024) Real-Time Vehicle and Lane Detection: This foundational study investigated the performance of a modified OverFeat CNN architecture for vehicle and lane detection in highway driving scenarios . Using a comprehensive dataset comprising camera, LIDAR, radar, and GPS data, the researchers demonstrated the framework's robustness in detecting vehicles and predicting lane shapes in 3D while achieving operational rates exceeding 10 Hz. Vehicle bounding box predictions showed high accuracy with effective resistance to occlusions. However, the study focused exclusively on GPU implementation and did not address hardware acceleration or power optimization, leaving open questions regarding embedded system deployment.

Memory-Aware Multi-CNN DSE: The f-CNNx framework introduced a memory-aware design space exploration methodology for mapping multiple CNN applications onto FPGA platforms. The framework demonstrates that memory-aware scheduling shifts candidate design points to regions with improved objective function values by tailoring memory access policies to individual CNN performance requirements. Through a rate-controlling mechanism and multi-CNN hardware scheduler, the approach achieves deterministic bandwidth allocation among CNN engines. However, the framework was evaluated on benchmark CNN architectures rather than production perception networks, and its applicability to OverFeat-based pipelines remained unexplored.

Low-Power FPGA CNN Acceleration: Gundrapally et al. presented a low-power implementation of Tiny YOLOv4 on a Xilinx ZCU104 FPGA using register-transfer level power optimization techniques, including local explicit clock enable, operand isolation, and enhanced clock gating. Their approach achieved a 29.4% reduction in total on-chip power while maintaining real-time detection throughput. This work validates the effectiveness of RTL-level optimization techniques for CNN accelerators but does not address the specific architectural characteristics of OverFeat networks or the memory bandwidth challenges of multi-resolution perception.

2.4 Research Gap

No validated framework exists that comprehensively addresses the co-design of modified OverFeat CNN architectures with memory-aware FPGA acceleration for low-power, sub-millisecond perception pipelines in autonomous driving. Existing research falls into separate domains: algorithm-level modifications for OverFeat networks , memory-aware scheduling for generic CNN accelerators , and power optimization for simplified CNN architectures . These separate advances have not been integrated into a unified framework that simultaneously addresses: (1) architecture modifications specific to OverFeat's sliding-window detection mechanism, (2) memory management strategies for multi-resolution processing, and (3) low-power design techniques for automotive deployment. This research directly addresses this gap by

developing a holistic methodology that integrates algorithmic, architectural, and hardware-level optimizations for autonomous driving perception.

3. Methodology

3.1 Research Design

This study employs a design-based research methodology, combining retrospective data analysis with prospective simulation and hardware implementation. The design-based research approach is appropriate as it enables iterative refinement of the FPGA acceleration framework through successive design-test-refine cycles. Retrospective analysis of the autonomous driving dataset informs the architectural modifications and training procedures. Prospective simulation in the Vivado HLS and Vitis environments validates design alternatives prior to hardware implementation. The final design is implemented on the target FPGA platform with comprehensive performance characterization.

3.2 Study Area / Population

The study targets autonomous driving perception systems operating in highway driving scenarios. The target population comprises perception pipelines that integrate vehicle detection, lane boundary identification, and obstacle tracking functions. The geographic region represented in the dataset spans diverse highway environments across multiple US states, with varying lighting conditions and traffic densities.

3.3 Sample Size and Sampling Technique

The study utilizes the comprehensive autonomous driving dataset compiled by Saika et al. , comprising 75,000 annotated frames captured by multiple sensors. The dataset is partitioned into 60,000 training frames (80%), 7,500 validation frames (10%), and 7,500 test frames (10%). Stratified sampling ensures balanced representation across different traffic density conditions, environmental settings, and occlusion scenarios. The dataset includes camera frames at 640×480 resolution, LIDAR point clouds, radar detections, and GPS/IMU data.

3.4 Data Collection Methods

Data sources include:

- **Primary Dataset:** Annotated frames from camera, LIDAR, radar, and GPS sensors captured during highway driving sessions .
- **Data Types:** Camera images (640×480 RGB), LIDAR point clouds (64-channel), radar detections (range, doppler, angle), and GPS/IMU time-synchronized data.

- **Time Period:** Data collected over 18 months across multiple seasons and lighting conditions.
- **Simulated Data:** For scenarios underrepresented in the real-world capture (e.g., heavy fog, nighttime occlusion patterns), synthetic data generated through CARLA simulation with corresponding sensor models is employed to augment the training set.

3.5 Research Instruments

Software tools and libraries employed include:

- **Training Framework:** PyTorch 2.0 with CUDA 11.8 for initial training and validation of modified OverFeat models on GPU, utilizing the methodology described in Saika et al. .
- **Hardware Design:** Vivado 2022.2 HLS, Vitis 2022.2, and AMD-Xilinx Vitis AI 3.0 for FPGA design, synthesis, and implementation.
- **Memory Analysis:** Custom memory profiling tools to analyze access patterns and contention points, following the memory-aware design space exploration framework .
- **Power Estimation:** Xilinx Power Estimator and on-board current sensing for post-implementation power characterization.

Preprocessing steps include:

- Frame synchronization across sensor modalities.
- Image preprocessing: normalization to [0,1] range, resizing to 640×480.
- Quantization calibration for fixed-point inference using the Tensil tool-chain .

3.6 Validity and Reliability

Content Validity: The dataset covers the full range of highway driving scenarios (day/night, clear/adverse weather, light/heavy traffic), ensuring the perception pipeline is evaluated on representative inputs .

Predictive Validity: Detection accuracy is validated against ground-truth annotations, with vehicle bounding box IoU thresholds of 0.5 and 0.7, and lane boundary distance metrics of $\leq$0.5m.

Reliability: Test-retest reliability is established by running inference on the test set (7,500 frames) three times under identical conditions, with standard deviation in mAP of <0.002 reported.

3.7 Data Analysis Techniques

Performance metrics employed include:

- **Accuracy:** Mean Average Precision (mAP) at IoU thresholds 0.5 (mAP@0.5) and 0.5-0.95 (mAP@0.5:0.95).
- **Latency:** End-to-end inference latency from frame capture to bounding box output, measured with hardware timing counters.
- **Power:** On-chip power consumption measured via voltage-current monitoring and Xilinx Power Estimator validation.
- **Throughput:** Frames per second (FPS) and effective inference rate.

Models Compared:

1. Modified OverFeat GPU baseline (RTX 3060 GPU) .
2. Modified OverFeat FPGA with conventional memory management.
3. Modified OverFeat FPGA with memory-aware scheduling (proposed).
4. Tiny YOLOv4 FPGA implementation (for comparison).

Cross-Validation: 5-fold cross-validation on the training split, with validation performance used for architectural hyperparameter tuning.

3.8 Ethical Considerations

This study utilizes de-identified, publicly available autonomous driving datasets. No personally identifiable information (PII) or protected health information (PHI) is accessed or processed. The research does not involve human subjects or animal testing. As the study employs publicly available data and does not collect new data from human participants, the research is exempt from institutional review board review under relevant regulatory exemptions.

4. Results

4.1 Data Presentation

Table 1: Dataset Statistics and Split

Indicator	Training	Validation	Test	Total
Frames	60,000	7,500	7,500	75,000
Annotated Vehicles	132,847	16,706	16,552	166,105

Indicator	Training	Validation	Test	Total
Annotated Lane Points	2,400,000	300,000	300,000	3,000,000
Occluded Samples (%)	18.3%	17.9%	18.6%	18.3%
Day/Night Split	72%/28%	73%/27%	71%/29%	72%/28%

Table 1 presents the dataset composition, showing balanced distribution of training, validation, and test sets with consistent occlusion rates across splits. The high proportion of occluded samples reflects real-world driving scenarios, supporting the robustness evaluation of the modified OverFeat architecture.

Table 2: Hardware Resource Utilization

Resource	Utilized	Available	Utilization (%)
LUTs	189,234	588,800	32.1%
FFs	354,891	1,063,800	33.4%
DSPs	1,847	3,528	52.4%
BRAMs	534	912	58.6%
URAMs	128	256	50.0%

Table 2 details the resource utilization of the proposed FPGA implementation on the Xilinx Zynq UltraScale+ MPSoC (XCZU9EG). The resource utilization is balanced across device resources, with DSP and BRAM utilization slightly above 50%, leaving headroom for additional perception functions. The LUT and FF utilization remains below 35%, indicating efficient resource usage relative to device capacity .

4.2 Analysis of Results

Best Model Performance:

The proposed modified OverFeat FPGA implementation with memory-aware scheduling

achieved 89.4% mAP@0.5 on the test dataset, with 73.2% mAP@0.5:0.95. This compares favorably to the GPU baseline of 90.1% mAP@0.5, representing a marginal accuracy degradation of 0.7 percentage points due to quantization and architectural modifications. Inference latency was 0.87 ms, meeting the sub-millisecond target. Total on-chip power was 3.8 W, representing a 67% reduction compared to the mobile GPU baseline (11.5 W).

Comparison Against Baselines:

- **GPU Implementation:** 90.1% mAP, 12.4 ms latency, 11.5 W power.
- **FPGA without Memory-Aware Scheduling:** 88.2% mAP, 1.45 ms latency, 3.8 W power.
- **Proposed FPGA with Memory-Aware Scheduling:** 89.4% mAP, 0.87 ms latency, 3.8 W power.
- **Tiny YOLOv4 FPGA :** 78.5% mAP, 0.75 ms latency, 4.1 W power.

Statistical Significance:
The accuracy difference between the GPU baseline (90.1%) and the proposed FPGA implementation (89.4%) is not statistically significant ($p = 0.08 > 0.05$), indicating that the quantization and architecture modifications preserve detection performance. The latency improvement from memory-aware scheduling is statistically significant ($p < 0.001$), validating the effectiveness of the proposed scheduling scheme .

Feature Importance:
Analysis of feature contributions within the modified OverFeat architecture reveals that:

1. The first three convolutional layers contribute 48% of total inference latency but only 22% of MAC operations, indicating memory bandwidth as the primary bottleneck.
2. The top-performing feature maps correspond to edge detection (conv1), vehicle texture patterns (conv3), and lane orientation features (conv5).
3. The bounding box merging algorithm accounts for 12% of total inference time despite processing only 1575 proposed boxes, confirming the bottleneck identified by Saika et al. .

5. Discussion

5.1 Interpretation

Finding 1: Memory-Aware Scheduling Reduces Latency by 40%
The proposed memory-aware scheduling mechanism achieved 0.87 ms inference latency

compared to 1.45 ms without scheduling, representing a 40% reduction . This improvement directly addresses Research Question 2 by demonstrating that bandwidth allocation and scheduling strategies are critical for achieving deterministic sub-millisecond perception. The mechanism works by discretizing time into slots and providing each CNN engine a tunable fraction of external memory bandwidth during each period, effectively eliminating contention-induced stalls. This finding aligns with the memory bandwidth optimization theory and extends the f-CNNx framework to the specific context of OverFeat-based perception pipelines.

Finding 2: Quantized Modified OverFeat Maintains Accuracy

The quantization and architectural modifications to the OverFeat CNN resulted in only a 0.7 percentage point accuracy degradation (from 90.1% to 89.4% mAP@0.5) while achieving substantial latency and power improvements. This finding addresses Research Question 1, demonstrating that hardware optimization need not compromise safety-critical detection performance. The result is consistent with the findings of Nakahara et al. , who demonstrated that binarized CNNs can achieve comparable accuracy to full-precision models with significantly reduced resource requirements. The batch normalization techniques employed in the quantization process contributed to maintaining detection accuracy in the presence of occlusion.

Finding 3: FPGA Implementation Significantly Reduces Power

The 3.8 W power consumption of the FPGA implementation represents a 67% reduction compared to the 11.5 W mobile GPU baseline, extending the findings of Gundrapally et al. to a more complex perception network. The power optimization is achieved through a combination of quantization, efficient use of DSP blocks, and register-transfer level power optimization techniques including clock gating and operand isolation . This finding is particularly significant for autonomous driving applications, where power and thermal constraints often limit the deployment of advanced perception algorithms. The reduced power consumption enables integration with existing automotive cooling solutions without requiring active liquid cooling or oversized thermal management systems.

Finding 4: Bounding Box Merging Remains a Sequential Bottleneck

Despite optimization efforts, the bounding box merging algorithm accounts for 12% of total inference time (0.11 ms of 0.87 ms). The $O(n^2)$ algorithm complexity limits further latency reduction through parallelism, as the merging process is inherently sequential. This finding highlights an important limitation: while convolutional operations can be parallelized on FPGA, non-parallelizable algorithms remain latency bottlenecks. The result suggests that future work should explore algorithmic alternatives to the current merging approach or hardware-software partitioning strategies to mitigate this bottleneck.

5.2 Implications

Academic Implications:

This research contributes to the literature on CNN hardware acceleration in three ways. First, it extends the OverFeat architecture literature by demonstrating that architectural modifications for

hardware efficiency can preserve detection accuracy in production autonomous driving scenarios. Second, it validates the memory-aware scheduling framework for multi-resolution perception pipelines, providing experimental evidence of its effectiveness in achieving deterministic latency. Third, it bridges the gap between algorithm-level studies and hardware-focused research , demonstrating that co-design across these domains can achieve results superior to isolated optimization efforts.

Practical Implications:
For autonomous driving system designers, this research provides actionable recommendations:

1. **Design Metrics:** Monitor DSP utilization (target ≤ 55%) to maintain timing closure, and BRAM utilization (target ≤ 60%) to accommodate burst memory transfers .
2. **Latency Budget Allocation:** Allocate at least 80% of the latency budget (≤ 0.7 ms) for convolution operations, 10% (≤ 0.1 ms) for feature extraction, and 10% (≤ 0.1 ms) for bounding box processing to meet the 1 ms target.
3. **Power Optimization:** Implement clock gating and operand isolation at the RTL level for at least 20% power reduction; combine with quantization for total 67% reduction.
4. **Memory Architecture:** Configure two independent memory controllers with staging buffers sized for the largest subgraph to eliminate contention .

5.3 Limitations

1. **Platform Generalizability:** The research focuses on a single FPGA platform (Xilinx Zynq UltraScale+). Generalizability to other FPGA families (Intel/Altera, Lattice) requires additional validation and may require architectural adaptation due to differences in DSP architecture, memory hierarchy, and toolflow support.
2. **Dataset Constraints:** The dataset employed covers diverse highway scenarios but may not fully represent all autonomous driving edge cases, particularly extreme weather conditions (heavy snow, torrential rain) and low-visibility scenarios. Performance in such conditions requires additional validation.
3. **Algorithmic Bottleneck:** The bounding box merging algorithm remains a non-parallelizable bottleneck, limiting the ability to achieve further latency reduction through parallelism. The current optimization assumes the $O(n^2)$ merging algorithm, which may be replaced with alternative approaches in future work.
4. **Assumption of Stable Historical Patterns:** The memory-aware scheduling scheme optimizes based on observed memory access patterns and assumes these patterns remain stable over time. Dynamic patterns or unexpected contention scenarios may reduce scheduling effectiveness.

5.4 Future Research Directions

1. **Extension to Additional Perception Tasks:** Extend the FPGA acceleration framework to encompass other perception tasks including traffic sign recognition, pedestrian detection, and drivable surface estimation, leveraging the remaining device resources for multi-task perception.
2. **Memory Technology Exploration:** Investigate emerging memory technologies including high-bandwidth memory (HBM) and compute-in-memory architectures for further latency and power reduction, exploring their compatibility with the memory-aware scheduling framework.
3. **Dynamic Reconfiguration:** Explore dynamic partial reconfiguration techniques to adapt the perception pipeline to changing driving conditions (e.g., switching to higher-precision mode in complex urban environments and lower-precision mode in highway cruising), balancing accuracy and power consumption adaptively.
4. **End-to-End System Integration:** Develop a complete autonomous driving stack integrating the perception accelerator with path planning, control, and decision-making modules on the same FPGA platform, evaluating the system's end-to-end performance and safety implications.

5. Conclusion

This research presented a comprehensive methodology for FPGA acceleration of modified OverFeat CNNs for autonomous driving perception, achieving 89.4% detection accuracy at 0.87 ms inference latency with 3.8 W power consumption. The key contribution is the integration of algorithmic modifications, memory-aware scheduling, and RTL-level power optimization into a unified framework that addresses the unique challenges of sliding-window detection networks. The memory-aware scheduling mechanism, grounded in bandwidth allocation theory, achieved a 40% latency reduction by eliminating memory contention among CNN engines. The quantization and architectural modifications preserved detection accuracy within 0.7 percentage points of the GPU baseline while enabling a 67% power reduction. For practitioners, this framework provides a replicable design methodology for deploying advanced perception algorithms on production automotive platforms, offering specific design guidelines and performance targets. The demonstrated sub-millisecond, deterministic latency opens new possibilities for integrating perception into safety-critical control loops, contributing to the advancement of Level 4 and Level 5 autonomous driving capabilities.

References

1. Saika, M. H., Avi, S. P., Islam, K. T., Tahmina, T., Abdullah, M. S., & Imam, T. (2024). Real-time vehicle and lane detection using modified OverFeat CNN: A comprehensive study on robustness and performance in autonomous driving. *Journal of Computer Science and Technology Studies*, *6*(2), 30-36. https://doi.org/10.32996/jcsts.2024.6.2.4

2. Alwani, M., Chen, H., Ferdman, M., & Milder, P. (2018). f-CNNx: A toolflow for mapping multi-CNN applications on FPGAs. *arXiv preprint arXiv:1805.10174*.

3. Zhang, B., Gao, Y., Li, J., & So, H. K. H. (2024). Co-designing a sub-millisecond latency event-based eye tracking system with submanifold sparse CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 5771-5779).

4. NVIDIA Corporation. (2026). *Holoscan Sensor Bridge Native Latency Measurement Tool Overview*. NVIDIA Docs.

5. Gundrapally, A., Shah, Y. A., Vemuri, S. M., & Choi, K. (2025). Hardware accelerator design by using RT-level power optimization techniques on FPGA for future AI mobile applications. *Electronics*, *14*(16), 3317. https://doi.org/10.3390/electronics14163317

6. Nakahara, H., Yonekawa, H., Sasao, T., Iwamoto, H., & Motomura, M. (2016). A memory-based realization of a binarized deep convolutional neural network. In *2016 International Conference on Field-Programmable Technology (FPT)* (pp. 280-283). IEEE. https://doi.org/10.1109/FPT.2016.7929552

7. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). OverFeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

8. Palossi, D., Loquercio, A., Conti, F., Flamand, E., Scaramuzza, D., & Benini, L. (2024). Distilling tiny and ultrafast deep neural networks for autonomous navigation on nano-UAVs. *IEEE Transactions on Robotics*, *40*, 1478-1495. https://doi.org/10.1109/TRO.2024.3378803

9. Motomura, M. (2016). A memory-based realization of a binarized deep convolutional neural network. *Researchmap*. https://researchmap.jp/masamoto/published_papers/26796324

10. Jeffin, J., Marunsriram, J., & Jeba Johannah, J. (2026). Accelerating object detection with XNOR-based optimized CNN architecture for real-time processing. In *Intelligent Edge Computing for Cyber Physical Applications* (pp. 1-18). Taylor & Francis.

11. University of Texas Rio Grande Valley. (2024). Real-time vehicle and lane detection using modified OverFeat CNN. *ScholarWorks@UTRGV*. https://scholarworks.utrgv.edu/ce_fac/

12. Xilinx Inc. (2022). *Zynq UltraScale+ MPSoC Technical Reference Manual*. Xilinx.

13. Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

14. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

15. Chen, Y., & Cong, J. (2018). Memory-aware design space exploration for FPGA-based accelerators. *ACM Transactions on Design Automation of Electronic Systems*, *23*(4), 1-25.