"Energy-Efficient FPGA-Based CNN Accelerator for Real-Time Edge Inferences"

Yashaswini M.[1], Prof. Shilpa B M.[2], Drakshayini K R.[3], K Shreya.[4], Sushmitha.[5]
[1,2,3,4,5]*Sri Krishna Institute of Technology, Bengaluru-560090*

***Abstract*—The increasing demand for real-time artificial intelligence applications has driven the need for efficient hardware acceleration of Convolutional Neural Networks (CNNs), particularly in edge computing environments. However, the high computational complexity and energy requirements of CNN models limit their deployment on resource-constrained platforms. Field-Programmable Gate Arrays (FPGAs) have emerged as a promising solution due to their reconfigurability and parallel processing capabilities. This paper presents a survey of recent FPGA-based CNN accelerators with a focus on architectural optimizations, dataflow scheduling, precision reduction, and reconfigurable designs. The study analyzes various approaches aimed at improving performance, reducing power consumption, and enhancing hardware utilization. The review highlights key techniques such as depthwise separable convolutions, adaptive dataflow mechanisms, runtime reconfiguration, and model compression strategies. Finally, the paper identifies existing limitations and discusses future directions for developing more efficient and scalable FPGA-based CNN accelerators for edge inference applications.**

***Index Terms*—FPGA, CNN Accelerator, Edge Computing, Deep Learning, Energy Efficiency, Reconfigurable Architecture**

I. INTRODUCTION

The rapid growth of deep learning applications has significantly increased the demand for high-performance computing systems capable of executing Convolutional Neural Networks (CNNs) efficiently. While Graphics Processing Units (GPUs) provide high computational power, they are often unsuitable for low-power and latency-sensitive edge devices. As a result, FPGA-based CNN accelerators have gained attention due to their ability to offer parallel computation, low energy consumption, and hardware customization.

Edge computing applications such as autonomous systems, healthcare monitoring, and intelligent surveillance require fast inference with minimal delay. However, CNN models are computationally intensive and require optimized hardware architectures for real-time execution. Therefore, researchers have focused on designing FPGA-based accelerators that improve efficiency through architectural optimization and resource-aware implementation.

This paper provides a structured review of recent FPGA-based CNN accelerator designs, focusing on their architectural strategies, optimization techniques, and performance improvements.

II. LITERATURE REVIEW

Recent advancements in edge computing have increased the demand for efficient hardware acceleration of Convolutional Neural Networks (CNNs). Due to high computational and memory requirements, FPGA-based accelerators have emerged as a promising solution for real-time inference in resource-constrained environments. Existing research primarily focuses on optimizing convolution operations, improving dataflow efficiency, and enhancing energy performance through architectural innovations.

Shen et al. [1] in 2025 proposed an efficient CNN accelerator for low-end FPGA platforms by optimizing depth wise separable convolutions and integrating squeeze-and-excite modules. The design reduces computational complexity and hardware resource usage while maintaining inference accuracy. Experimental results show significant performance improvement over CPU-based implementations and reduced DSP utilization, making it suitable for edge applications.

Dharun Kanna et al. [2] in 2025 implemented a CNN accelerator using Vivado HLS on a PYNQ FPGA board. The design employs PIPELINE and UNROLL optimizations along with efficient on-chip memory management using BLOCK pragmas. The accelerator achieves significant speedup over software execution, demonstrating improved latency, throughput, and energy efficiency for edge inference tasks.

Hung K. Nguyen et al. [3] 2024 developed an FPGA-based CNN accelerator for COVID-19 detection using chest X-ray images. The architecture includes convolution and max-pooling units for efficient feature extraction. The results confirm improved processing speed and reduced energy consumption, highlighting its applicability in medical edge systems.

Manca et al. [4] in 2025 presented a comparative study of FPGA CNN accelerator design flows, evaluating frameworks such as HLS4ML, FINN, and Vitis AI. Their work introduces runtime reconfigurable architectures using Multi-Dataflow Composer, improving adaptability while highlighting trade-offs between performance and design complexity.

Fata and Elmannai [5] in 2026 proposed a low-cost FPGA-based accelerator for real-time YOLO object detection using pruning, quantization-aware training, and optimized memory mapping. The design significantly reduces model size while maintaining accuracy and achieves low power consumption with high detection performance on edge devices.

Xu, Luo, and Sun [6] in 2024 introduced a full-precision reconfigurable CNN accelerator using vector dot-product unification and dynamic partial reconfiguration. The architecture improves resource utilization and power efficiency while achieving high throughput across AlexNet and VGG16 models, demonstrating strong scalability and performance benefits.

III. RESEARCH GAP

From the reviewed literature, several limitations are identified. Most existing FPGA-based CNN accelerators focus on specific optimization techniques such as quantization, pruning, or dataflow scheduling, but lack a unified architecture combining multiple optimizations. Additionally, many designs are application-specific and do not generalize well across different CNN models. Limited research addresses the trade-off between full-precision computation and energy efficiency in reconfigurable architectures. Furthermore, integration of adaptive runtime optimization with hardware-level flexibility remains an open challenge for achieving scalable edge AI systems.Furthermore, most studies emphasize performance improvements without providing a holistic evaluation of power consumption, scalability, and flexibility. These gaps indicate the need for energy-aware, adaptable, and communication-efficient FPGA-based CNN accelerator designs suitable for next-generation edge inference systems.

IV. FUTURE SCOPE

Future FPGA-based CNN accelerator research can focus on hybrid architectures combining quantization, pruning, and adaptive scheduling techniques. The integration of dynamic reconfiguration with intelligent workload management can further improve efficiency. Additionally, exploring support for emerging deep learning models and transformer-based architectures will enhance applicability. Low-power design techniques combined with AI-driven hardware optimization can significantly improve edge inference performance in next-generation systems.

V. CONCLUSION

This paper presented a structured survey of FPGA-based CNN accelerators designed for edge inference applications. Various architectural techniques such as optimized convolution operations, adaptive dataflow scheduling, runtime reconfiguration, and hardware-efficient implementations were analyzed. The study shows that FPGA-based solutions provide significant advantages in terms of energy efficiency, flexibility, and real-time performance. However, challenges such as limited generalization, optimization trade-offs, and integration complexity still remain. Addressing these issues will be essential for developing next-generation intelligent edge computing systems.

REFERENCES

[1] J. Shen, X. Cheng, X. Yang, L. Zhang, W. Cheng, and Y. Lin, "Efficient CNN accelerator based on low-end FPGA with optimized depthwise separable convolution and squeeze-

and-excite modules," *Artif. Intell. J.*, vol. 58, no. 3, pp. 201–213, 2025.

[2] N. D. Kanna, G. S. Sadasivam, and P. Saravanan, "CNN accelerator for edge inference," in *Proc. IEEE Int. Conf. VLSI Design*, 2025, pp. 112–117.

[3] H. K. Nguyen, H.-T. Duong, and D.-T. Trang, "A high-performance CNN hardware accelerator for COVID-19 feature extraction from X-ray images," *IEEE Sensors J.*, vol. 24, no. 7, pp. 2239–2249, 2024.

[4] F. Manca, F. Ratto, C. Rubattu, L. Raffo, and F. Palumbo, "Adaptive CNN acceleration on FPGAs: Closing the gap with state-of-the-art solutions," *IEEE Access*, vol. 13, pp. 45621–45635, 2025.

[5] J. S. Fata and W. M. Elmannai, "Low-cost FPGA-enhanced CNN accelerator for real-time YOLO object detection and classification," vol. 14, 2026.

[6] Y. Xu, J. Luo, and W. Sun, "Flare: An FPGA-based full precision low-power CNN accelerator with reconfigurable structure," *Sensors*, vol. 24, Art. no. 2239, 2024.