# A Sustainable FPGA Based Edge Computing Platform for Computer Vision Applications

xxx
*xxx*
*xxx*
xxx
xxx

xxx
*xxx*
*xxx*
xxx
xxx

***Abstract*—Edge Computing has shown to be able to provide energy efficient and sustainable services with low latency because of proximity to the users. Edge Computing has been widely utilized for task executions in Smart Health and Smart Manufacturing. Computer Vision techniques have been broadly applied in the areas of Smart Health and Smart Manufacturing. Deep Learning has proved to be promising in deploying Computer Vision applications. However, traditional computing platforms with Central Processing Unit (CPU) and Graphic Processing Unit (GPU) fall short in completing computation intensive tasks of Deep Learning, regarding energy efficiency. In this study, Field Programmable Gate Array (FPGA), because of its intrinsic hardware reconfigurable features, works as Edge Computing device to perform computation intensive Deep Learning tasks. The experiment result shows that, comparing with traditional Central Processing Unit (CPU) and Graphic Processing Unit (GPU) based computation platforms, the developed FPGA based platform is able to get highest energy saving while still being able to complete computing jobs in a fast and accurate manner.**

***Index Terms*—Edge Computing, FPGA, Deep Learning, Smart Health, Smart Manufacturing**

## I. Introduction

Edge Computing paradigm targets at providing services to users in a fast and low energy cost manner [1]. It has been intensively used in different areas because of its physical locating close to the data sources and being able to provide immediate responses for critical tasks. As the techniques of Artificial Intelligence (AI), Internet of Things (IoT) and Wearable Computing develops, Smart Health gradually becomes a promising sector which absorbs different kinds of technology to transform healthcare into more pervasive and accessible services for patients. It proposes intelligent, efficient and affordable services for users and patients. Smart Manufacturing incorporates the technology of AI, IoT and digital twins into factories to enable data-driven and automation in production and fabrication procedures. It can help improve efficiency, provide diagnostics for equipment failure, inspect defects in manufacturing, and maintain a sustainable fabrication process to adapt to an accelerated production iteration cycle. Computer Vision techniques have been widely used in Smart Health and Smart Manufacturing. Deep Learning has been utilized in different Computer Vision applications of Smart Health and Smart Manufacturing [2]–[7]. However, traditional computing systems, even with powerful Central Processing Units (CPUs) and Graphic Processing Units (GPUs), are still having concerns with huge energy consumptions for computation intensive tasks of Deep Learning [2]–[4], [7]. Field Programmable Gate Array (FPGA) has been proved to be an efficient way of deploying computation intensive deep learning tasks [8], [9]. FPGA based hardware architecture has been implemented for sustainable biomedical health applications [10]. FPGA has been deployed for defect inspection in Additive Manufacturing (AM) [11]. In their work, FPGA has been proved to be able to perform real-time inspection of defects for AM with high accuracy.

Yadav *et al.* [12] present an FPGA based framework to implement medical image processing for Malaria and Pneumonia detection. The study explores the strategies to optimize convolution calculation procedures, improve computation execution efficiency and expand memory accessibility. The experiment result shows FPGA based platform achieves remarkable energy savings compared to other platforms. Hsu *et al.* [13] develop an FPGA based system to obtain heart rate remotely. The Photoplethysmography (PPG) based video is recorded for heart beat extraction. The implemented FPGA system exhibits outstanding performance in terms of computation time and accuracy. FPGA can be an important component in Edge Computing to provide healthcare services. FPGA based Edge Computing systems have been proposed to serve for different healthcare scenarios [14].

Khan *et al.* [9] summarize the usage of FPGA in distributed digital manufacturing industry. The manufacturing processes, which could incorporate FPGA, include Additive Manufacturing, Milling and Grinding processes. FPGA has also been used for motion control in industrial systems [15]. FPGA based computer vision systems have been used to identify wafer and die positions in an accurate and accelerated manner. FPGA can be an infrastructure device in Edge Computing to assist industrial manufacturing and quality control. An FPGA based Edge Computing platform has been proposed for Photovoltaic (PV) panel defect detection [16].

In this study, we implemented an Edge Computing platform, using off-the-shelf FPGA, which is able to achieve outstanding energy savings, compared to CPU/GPU based computing paradigms while still maintaining high processing speed and accuracy. We test the FPGA computing platforms

on different Computer Vision datasets. For Computer Vision applications in Smart Health scenarios, we validate our system performance on Breast Cancer image datasets. For Computer Vision applications in Smart Manufacturing scenarios, we test our method on metal defect image datasets. In these scenarios, the experiment results exhibit lower energy consumption of FPGA platform, compared to CPU/GPU based platforms while still keeping satisfying accuracy and timing performance. All these findings show that FPGA involved systems can be significant components for Edge Computing infrastructure and play important roles for Smart Health and Smart Manufacturing.

## II. Method and Experiment

In our implementation, we use Convolutional Neural Networks (CNN) as Deep Learning method for image classification. When deploying CNN over FPGA, there are some customization work which needs to be done. We refer to the work [17] to perform binarization for weights and activations. Binarization function below has been used to convert weights and activations from floating point to binary values:

$$z^b = \text{Sign}(z) = \begin{cases} +1 & \text{if } z > 0, \\ -1 & \text{otherwise}, \end{cases}$$

$z^b$ the binarized variable (weight or activation) and $z$ is the real-valued variable.

ResNet is a representative CNN architecture [18]. We use the binarized ResNet as CNN model to be implemented over all the three computing platforms: CPU, GPU and FPGA. The depth of the ResNet is set to 18. The CPU we use in the experiment is Intel Xeon CPU. The GPU we use in the experiment is the NVIDIA T4 GPU. The FPGA board used in the experiment is the PYNQ Z2. It is equipped with AMD Zynq-7000 System-on-Chips (SoC). We use FINN framework for Deep Learning model loading and quantization, to enable dataflow over FPGA [19]. We use AMD Vivado High-Level Synthesis (HLS) for FPGA design. As showed in Figure 1, PyTorch is used to train the Deep Learning model. ONNX is used to export the model. FINN framework helps load the model and perform the quantization. Xilinx Vivado generates the HLS accelerator, configures the dataflow architecture, and builds FPGA IP blocks. The bitfile is finally generated and deployed over FPGA.

As showed in the Figure 2, the ONNX format of model includes information of graph structure (topology), operator types, attributes (metadata) and weights (parameters). Graph structure (topology) provides records of nodes (which can be operators such as Convolution, Add and ReLU) and edges (which can be tensor connections between nodes). Each node has an operator type, such as Convolution, Matrix Multiplication, ReLU, etc. Attributes (metadata) include information about kernel size, stride, padding, etc. Weights (parameters) are stored as tensors (initializers).

Within the FINN design flow and hardware library, the Matrix Vector Threshold Unit (MVTU) is the key component of the generated FPGA accelerator. A large volume of

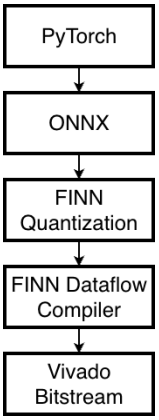


Fig. 1. Workflow to Deploy CNN to FPGA

computing operations in Deep Learning can be divided into groups of matrix-vector operations followed by thresholding. The MVTU can be used to build both fully-connected layers and convolutional layers. The architecture of the MVTU is shown in Figure 3. The MVTU includes multiple Processing Elements (PEs), each of which is equipped with a number of SIMD lanes. Inside MVTU, a PE can work as a neuron and an SIMD can work as synapse to simulate the functionality of neural network. When the quantized ResNet model is loaded into FINN, in order to make full use of the restricted resources of FPGA, FINN performs a special step called $Folding$ to map the ResNet neural network model into limited number of PEs as neurons and SIMDs as synapses, for hardware implementation. As the number of PEs and SIMDs increases, the extent of parallelism can be definitely enhanced to provide faster processing speed for FPGA. But it also means more resources dedicated, such as Look-Up Tables (LUTs) and Block RAM (BRAM), which would easily reach the limitations of capacity over the chip. The FINN library also develops different optimization strategies for specific computation units, such as Sliding Window Unit (SWU) for Convolution layer and the Pooling Unit (PU) for Max-pooling layer implementations.

We testify our FPGA platform in Smart Health and Smart Manufacturing scenarios:

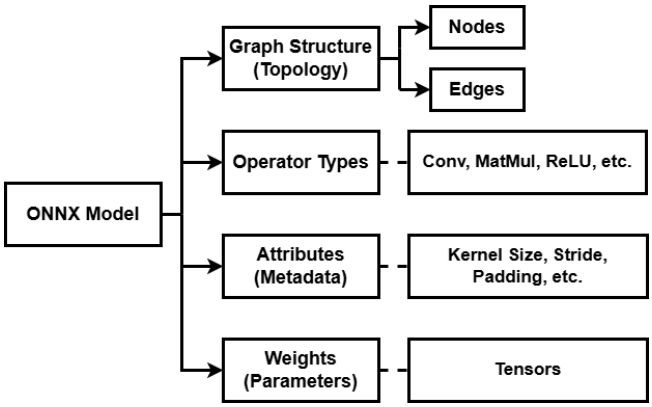


Fig. 2. ONNX Model

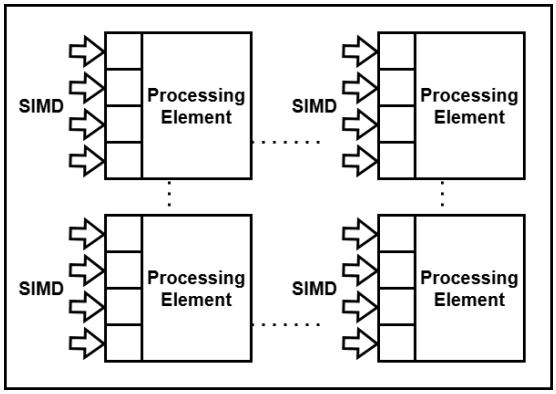


Fig. 3. Architecture of MVTU

- **Smart Health:** We test our method over the Breast Cancer image dataset [20]. The images have been labeled benign or malignant. The dataset currently contains four histological distinct types of benign breast tumors: Adenosis (A), Fibroadenoma (F), Phyllodes Tumor (PT), and Tubular Adenona (TA); and four malignant tumors (Breast Cancer): Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC) and Papillary Carcinoma (PC), as showed in Figure 4, 5 and 6.

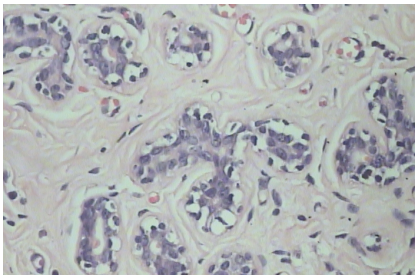

Fig. 4. Tubular Adenona (TA)

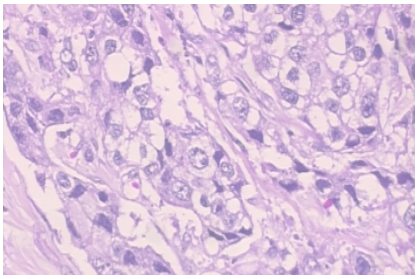

Fig. 5. Lobular Carcinoma (LC)

- **Smart Manufacturing:** We use metal surface defect image datasets GC10-DET [21]. GC10-DET is the surface defect dataset collected in a real industry. It contains ten types of surface defects, i.e., Punching (Pu), Weld line (Wl), Crescent gap (Cg), Water spot (Ws), Oil spot (Os), Silk spot (Ss), Inclusion (In), Rolled pit (Rp), Crease (Cr), Waist folding (Wf), as showed in Figure 7, 8, 9 and 10. The collected defects are on the surface of the steel sheet. The dataset includes 3570 gray-scale images.

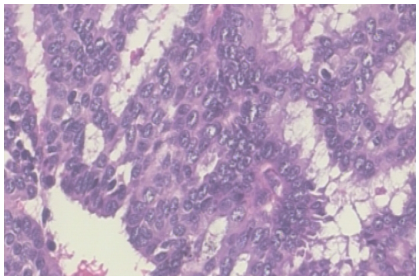

Fig. 6. Papillary Carcinoma (PC)

Fig. 7. Welding Line

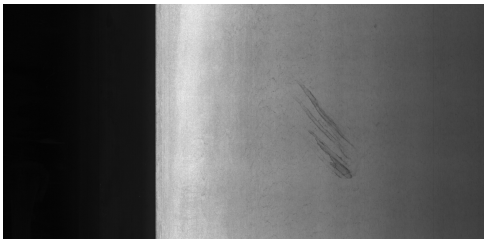

Fig. 8. Water Spot

Fig. 9. Waist Folding

Fig. 10. Crescent Gap

We measure the performance of Deep Learning models deployed over different platforms, in terms of processing accuracy (percentage), latency (millisecond) and energy consumption (Watt). Nvidia $smi$ command has been used to record the power consumption of GPU platform. $Perf$ has been used for CPU power consumption monitoring. PMBus interface is used to monitor the FPGA power supply rails.

In Table I, we present the testing result for the Breast Cancer dataset. The dataset contains 2013 images: 1390 positive (malignant) images and 623 negative (benign) images. When feeding to ResNet for classification, the images have been resized to 224x224.

TABLE I
COMPARISON RESULT FOR DEEP LEARNING OVER DIFFERENT PLATFORMS (SMART HEALTH)

| Methods | Accuracy (%) | Latency (ms) | Energy Consumption (W) |
|---|---|---|---|
| GPU | 93.71% | 5 ms | 224 W |
| CPU | 93.71% | 78 ms | 57 W |
| FPGA | 91.65% | 237 ms | 3 W |

In Table II, we present the testing result for manufacturing defect dataset GC10-DET. When feeding to ResNet, the images have been resized to 224x224.

TABLE II
COMPARISON RESULT FOR DEEP LEARNING OVER DIFFERENT PLATFORMS (SMART MANUFACTURING)

| Methods | Accuracy (%) | Latency (ms) | Energy Consumption (W) |
|---|---|---|---|
| GPU | 95.47% | 4 ms | 185 W |
| CPU | 95.47% | 63 ms | 49 W |
| FPGA | 92.68% | 126 ms | 2 W |

The accuracy degradation of FPGA, compared to GPU and CPU, is because of our customized hardware design, which perform the simplification and approximation. The result shows that FPGA based implementations could reach more energy saving while still maintaining high accuracy and fast speed.

## III. CONCLUSION

In this work, we show that Computer Vision techniques can be broadly used for pattern recognition and intelligent decision makings. FPGA is an affordable and sustainable device to implement Deep Learning method for Computer Vision tasks. We study deployment of the FPGA based platform for Smart Health and Smart Manufacturing. The result shows that FPGA based implementation methods could achieve lower energy cost while still being able to reach high accuracy and speed, for both clinical settings and industrial manufacturing. This study paves the way for future work, which integrates FPGA into IoT infrastructures to support more robust and versatile systems. In addition, more advanced Deep Learning models can be explored for implementation over FPGA to improve image classification performance. More sophisticated FPGA hardware design is worth investigating to further elaborate timing performance and power savings on the chips.

## REFERENCES


[1] C. Xu, S. Jiang, G. Luo, G. Sun, N. An, G. Huang, and X. Liu, "The case for fpga-based edge computing," *IEEE Transactions on Mobile Computing*, vol. 21, no. 7, pp. 2610–2619, 2020.

[2] B. Standfield, W.-B. Chen, Y. Wang, Y. Lu, A. F. Abdelzaher, X. Wang, and X.-G. Yang, "Using Convolutional Neural Networks to Detect and Extract Retinal Blood Vessels in Fundoscopic Images," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019, pp. 222–227.

[3] X. Wang, Y. Lu, and W.-B. Chen, "Promote Retinal Lesion Detection for Diabetic Retinopathy Stage Classification," in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2020, pp. 31–34.

[4] X. Wang, Y. Lu, W.-B. Chen, and D. Baker, "A deep learning based method for vitreous hemorrhage recognition in fundoscopic images," in *2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2023, pp. 233–234.

[5] K. Li, W.-B. Chen, Y. Lu, X. Wang, and H. Gao, "Automated recognition of optic disc and blood vessels in diabetic fundoscopy images using real-time image analysis," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2024, pp. 579–585.

[6] Y. Lu, W.-B. Chen, X. Wang, Z. Ailsworth, M. Tsui, H. Al-Ghaib, and B. Zimmerman, "Deep Learning-Based Models for Porosity Measurement in Thermal Barrier Coating Images," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 11, no. 3, pp. 20–35, 2020.

[7] X. Wang, J. Lowery, Y. Lu, W.-B. Chen, S. Zhang, and Z. Wu, "Embracing deep learning for crack segmentation in sem images of metal additive manufacturing," in *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2022, pp. 18–19.

[8] A. K. Jameil and H. Al-Raweshidy, "Efficient cnn architecture on fpga using high level module for healthcare devices," *IEEE Access*, vol. 10, pp. 60 486–60 495, 2022.

[9] L. Khan, S. P. Isanaka, and F. Liou, "Fpga-based sensors for distributed digital manufacturing systems: A state-of-the-art review," *Sensors*, vol. 24, no. 23, p. 7709, 2024.

[10] M. R. Azghadi, C. Lammie, J. K. Eshraghian, M. Payvand, E. Donati, B. Linares-Barranco, and G. Indiveri, "Hardware implementation of deep network accelerators towards healthcare and biomedical applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 6, pp. 1138–1159, 2020.

[11] Y. Luo and Y. Chen, "Fpga-based acceleration on additive manufacturing defects inspection," *Sensors*, vol. 21, no. 6, p. 2123, 2021.

[12] A. Yadav, V. K. Gupta, and B. Kumar, "Fpga-based medical image processing using hardware-software co-design approach," *IEEE transactions on biomedical circuits and systems*, 2025.

[13] J.-Y. Hsu, T.-Y. Jiang, and P. C.-P. Chao, "A fast fpga hardware accelerator for remote heart rate detection based on rgb vision," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 18, no. 3, pp. 592–607, 2024.

[14] M. Hartmann, U. S. Hashmi, and A. Imran, "Edge computing in smart health care systems: Review, challenges, and research directions," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 3, p. e3710, 2022.

[15] C. Rubattu, A. Ledda, F. Ratto, C. Jugade, D. Goswami, and F. Palumbo, "Integrating fpga-based acceleration in industrial motion control system," *IEEE Open Journal of the Industrial Electronics Society*, 2025.

[16] J. Ye, Y. Liu, H. Chen, C. Wang, Y. Zhou, L. Yang, and W. Zhang, "Edge computing accelerator for real-time defect detection of photovoltaic panel on lightweight fpgas," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[17] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] Y. Umuroglu, N. J. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre, and K. Vissers, "Finn: A framework for fast, scalable binarized neural network inference," in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '17. ACM, 2017, pp. 65–74.

[20] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *Ieee transactions on biomedical engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[21] X. Lv, F. Duan, J.-j. Jiang, X. Fu, and L. Gan, "Deep metallic surface defect detection: The new benchmark and detection network," *Sensors*, vol. 20, no. 6, p. 1562, 2020.