2021 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-7281-9201-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISCAS51556.2021.9401799

Developing a Miniature Energy-Harvesting-Powered Edge Device with Multi-Exit Neural Network

Yuyang Li, Yawen Wu, Xincheng Zhang, Ehab Hamed, Jingtong Hu, Inhee Lee

University of Pittsburgh, Pittsburgh, PA

{yul230, yawen.wu, xiz193, eah115, jthu, inhee.lee}@pitt.edu

***Abstract*—This paper describes a miniature edge device that performs neural network inference with different exit options depending on available energy. In addition to the main-exit path, it provides an alternative, early-exit path that requires less computation and thus increase the number of inference operations for given energy. To compensate its degraded accuracy, the proposed device provides entropy as a confidence level for the early exit. The network is implemented with a custom low-power 180 nm CMOS processor chip and a 90 nm embedded flash memory chip and tested by images from CIFAR-10 dataset. The measurement results show the proposed neural network reduces processing time and thus energy consumption by 41.3% compared with the main-exit only method while sacrificing its accuracy from 69.5% to 66.0%.**

Keywords*—Energy harvesting, neural network, multi-exit, miniature system.

I. Introduction

Rapid evolution in computing systems has continuously reduced their form factor over 70 years from a room-size mainframe to a miniature Internet-of-Things (IoT) device with centimeter/millimeter-scale size [1]. The miniature system becomes an attractive monitoring solution in a variety of applications including ecological, biomedical, security, and infrastructure [2]–[5]. Their tiny system size enables to measure important parameters with minimally affecting normal characteristics of a target object or even creates an unprecedented monitoring approach in a space-limited application (e.g., intraocular pressure sensor in an eye [6]).

However, there are still substantial challenges in miniature system design, including limited memory size and battery capacity. For applications where raw data size is significant, it is not efficient to store the original data into a small memory in the miniature system. For instance, a small image with 32 × 32 pixels, RGB color channel, and 8-bit resolution requires 3 kB (e.g., CIFAR-10 dataset [7]). A 128 kB flash memory [8], designed for a miniature system, can store only 42 images. To increase the number of data to be stored, it is more efficient to extract key information from the raw images and save only the significantly reduced data into the memory. For example, Convolutional Neural Network (CNN), one of recently highlighted machine learning (ML) techniques, categorizes CIFAR-10 dataset to 10 classes so that the final output can be stored in 4 bits [9]. A CNN with 70 kB coefficients enables to store both the coefficients and 116k inferenced results in the memory.

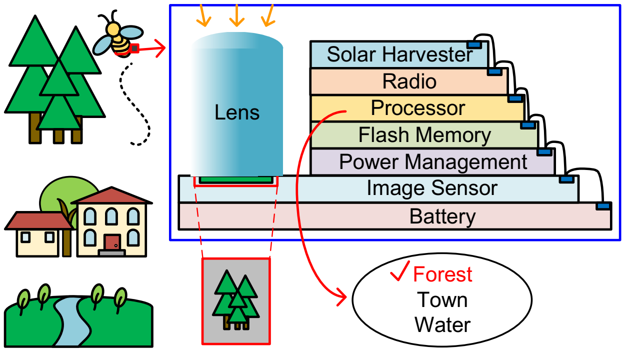


Fig. 1 Proposed system and an example application.

The ML technique performs intensive computation and consumes considerable energy. Thus, implementing it in a miniature system encounters an issue with limited available energy. The total system size budget restricts physical battery form factor and thus battery capacity. For example, a thin-film battery with 5.7 mm × 6.1 mm stores only electrical charge of 50 µAh or energy of 684 mJ [10]. A low-power TI MSP430 processor can discharge the battery in 4.2 days just by its standby power of 0.5 µA [11]. Moreover, it is very difficult to replace a battery once the millimeter-scale system is fully encapsulated for physical protection [12]. Hence, a miniature system usually employs energy harvesting technique to recharge the battery using environment energy and thus extends the system lifespan [12]–[14]. However, it does not always guarantee stable energy supply for extensive ML computation since the environment energy source is typically weak and intermittent [15].

In this paper, we present a design example of CNN implementation for a millimeter-scale energy-harvesting-powered edge device. It employs two exit options such as the main and early exits and chooses one according to available energy in a battery. The early-exit path provides faster processing time but less accuracy than the main-exit path. The CNN is implemented with a custom low-power processor chip [1] and a 128kB flash memory [8], previously fabricated in 180 nm CMOS and 90 nm embedded flash technology, respectively. The proposed neural network reduces processing time and thus energy consumption by 41.3% compared with the main-exit only method while losing the accuracy from 69.5% to 66.0%.

978-1-7281-9201-7/21/$31.00 ©2021 IEEE

II. TARGETED SYSTEM OVERVIEW

Table I. Performance summary of the state-of-the-art low-power CISs.

Parameters	JSSC 2019 [18]	ISCAS 2020 [19]	ISSCC 2019 [20]	VLSIC 2020 [21]
Pixel resolution	320 ×320	4240 ×3216	792 ×528	640 ×480
Frame rate	5 fps	30 fps	5.6 fps	15 fps
ADC resolution	10 bits	10 bits	10 bits	10 bits
RGB/Mono.	RGB	RGB	RGB	Mono.
Energy/frame /pixel/bit	9.1 pJ	8.1 pJ	5.6 pJ	5.6 pJ
Estimated energy for an image[A]	224 nJ	199 nJ	138 nJ	138 nJ
Estimated energy for an image[B]	286 μJ	255 μJ	176 μJ	176 μJ

A: 32 × 32, RGB, and 8 bits, B: 1024 × 1024, RGB, and 10 bits

Fig. 1 shows the proposed system for an example application where a miniature smart imager is attached to a small animal (e.g., insects [16], [17]) to study its living environment. The system takes its surrounding pictures, extract their key information, and save the inference results (e.g., forest, water, town, etc.) in a memory. Once the animal comes near a gateway, the collected data will be remotely transmitted. The miniature imager mainly consists of image sensing, wireless communication, and data processing & storage parts.

Table 1 summarizes the state-of-the-art low-power CMOS image sensors (CISs) for the image sensing part [18]–[21]. Power consumption of CISs has been reduced for portable device and becomes small enough to be powered by energy harvesters [18], [22]–[25]. The prior works have been optimized for different pixel resolution, RGB/monochrome, and ADC resolution. In the target application, relatively small images will be periodically captured rather than continuous streaming. Thus, performances reported for low-performance setting (e.g., low pixel resolution, slow flame rate, etc.) are used for comparison. All the CISs consumes small energy so that they do not rapidly discharge a millimeter-scale battery (e.g., 684 mJ [10]) even considering images of 1M pixels, RGB, and 10 bits.

For wireless communication, [26] demonstrated an energy-efficient radio transceiver designed for a miniature system. It consumes the average power of 60.6 μW at the maximum data rate of 30.3 kbps in transmit mode while consuming 1.85 mW at the maximum data rate of 62.5 kbps in receive mode. It costs 790 ms and 47.9 μJ to transmit a 32×32- pixel, RGB, 8-bit image while requiring 16.9 minutes and 61.4 mJ for a 1M-pixel, RGB, 10-bit image. For the latter, only 11 images can be transmitted by the 684 mJ battery. Thus, data size to be transmitted must be reduced [27].

Michigan Micro Mote (M^3) platform [1] has been used to develop various millimeter-scale systems with different sensing modality [28]–[33], including imagers [4], [34]. The platform consists of multiple chips stacked in a vertical way as shown in Fig. 1. The structure provides the maximum functionality (or silicon area for integrated circuits) per unit volume. The modular platform enables easy system development for different applications. The layers are connected by bonding wires that form low-power data bus [35] and deliver regulated supply voltages. The proposed system includes chips such as a solar energy harvester [12], a processor [1], a flash memory [8], a radio transceiver [26], a power management unit [36], and a battery [10]. Here, the processor and flash memory manage the system operation, extract key information from the images by a neural network for data size reduction, and storage its required coefficients and inference results. The process includes a commercial ARM Cortex-M0 processor and a custom low-power 16 kB SRAM, fabricated in 180 nm CMOS process [1]. The 128 kB embedded flash memory is designed with custom low-power peripheral circuits and fabricated in 90nm NOR flash technology [8]. This non-volatile memory enables the system keep the stored inference results and neural-network coefficients even when the system loses power under extremely low energy condition. Once the battery is recharged by the energy harvester, the system operation is restored by copying its program code from the flash memory back to the SRAM memory of the processor chip [8].

III. PROPOSED NEURAL NETWORK INFERENCE

In the proposed miniature system with limited memory capacity, a neural network helps to reduce output data size and enables to store essential information from all the images. However, the neural network is typically resource-hungry so that prior networks cannot be directly applied to the target system without proper adjustment. For instance, a small MobileNetV2 has >3.4M coefficients and requires computing 300M multiply-accumulate (MAC) operations [37]. Only to complete the MAC operations, a microprocessor operating at 1 MHz with 16 kB memory needs to perform 208 times of data exchange with the external data storage, taking 5 minutes. The high energy consumption restricts its usage in the proposed system with limited available energy.

To overcome this issue, [38] proposes to exploit compression on neural networks and break task-based abstraction for a system powered by intermittent energy, but it requires multiple power cycles to accomplish an inference. As the harvested energy are not predictable nor stable, the inference can be postponed for long time until available energy reaches a threshold again, and thus it blocks the following operations. Instead, we choose to use a multi-exit approach proposed in [39] and [40]. In addition to the original, main network, it additionally has a branch with shallower network and provides a choice to save processing time and thus energy consumption by sacrificing its inference accuracy. [15] applies the early-exit approach to an energy-harvesting-powered device.

Fig. 2 shows the proposed CNN designed for the target system based on [15], which is designed to process images compatible with CIFAR-10 dataset. It mainly consists of 3 blocks such as the shared, early-exit, and main-exit parts. The input is an image with 32×32 pixels, RGB channels, and 8 bits. Each convolutional block includes a convolutional layer with filter size 5×5, a max-polling layer, and a Rectified Linear Unit (ReLU). When available energy is lower than a threshold (BAT_{TH}), it performs the shallower network using the early-exit path. To compensate its reduced accuracy, it computes entropy as a confidence level. If the entropy level is lower than a threshold (ENT_{TH}), the inference result is stored in the memory.

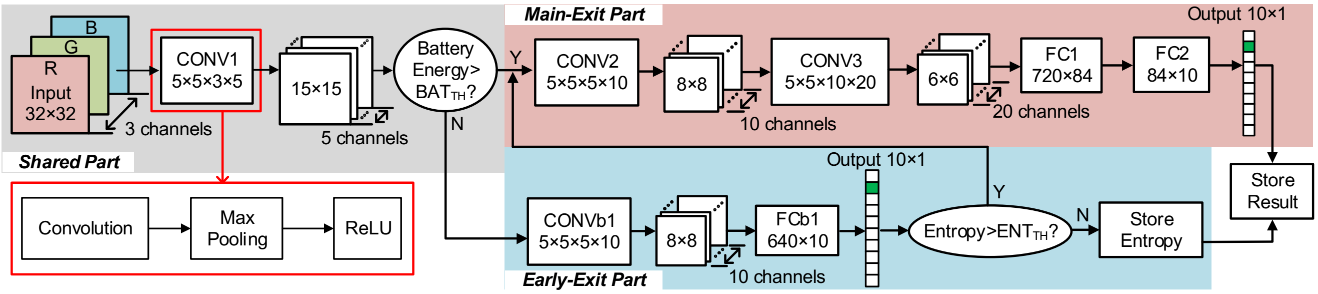


Fig. 2. Proposed CNN structure.

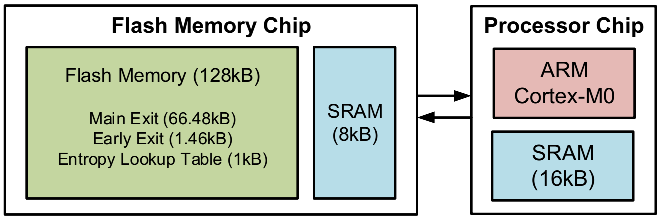
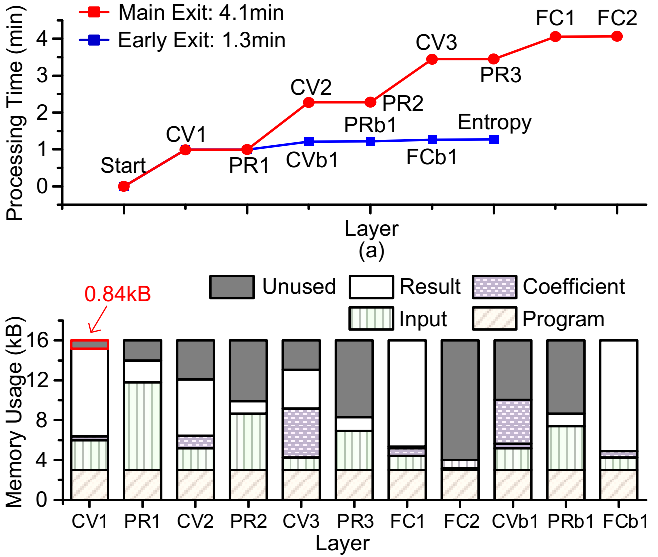


Fig. 3. Implemented CNN with a processor and flash memory chips.



(a)

(b)

Fig. 4. Simulated CNN layer. (a) Processing time. (b) Memory usage.

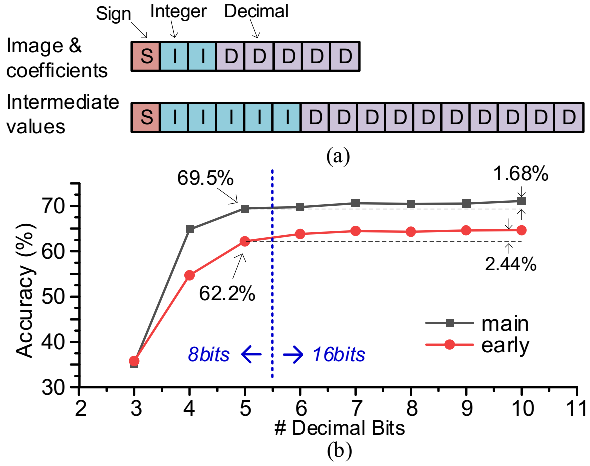
Otherwise, the result is aborted, and the main exit is executed instead. Here, the entropy is expressed as:

$$p_i = e^{output_i} / \sum e^{output_i}. \qquad (1)$$

$$Entropy = -1 \cdot \sum p_i \cdot \log_2 p_i. \qquad (2)$$

IV. Implementation of the Neural Network

The proposed neural network is implemented with custom processor and flash memory chips as shown in Fig .3. They are previously designed for low-power miniature systems and fabricated in 180 nm CMOS and 90 nm embedded flash technology, respectively [1], [8]. The processor chip includes an ARM Cortex-M0 processor and a 16 kB SRAM memory. The flash memory chip includes a 128 kB flash memory for data storage and also an 8 kB SRAM memory to transfer data between the SRAM of the processor and the flash memory. The two chips communicate through a low-power bus [35]. The processor chip receives image data and run the neural network by loading coefficients required for each layer from the flash memory chip. The coefficients include the weights and bias for the convolution and fully connected layers and the lookup table for entropy calculation. They are stored in the flash memory, taking 68 kB. The weights and bias are trained through Pytorch framework and then quantized to 8-bit fixed-point values. Each layer outputs 16-bit fixed-point results.



(a)

(b)

Fig. 5. Simulated CNN for 10k images with different bit allocations. (a) Bit field of the input and intermediate data. (b) Accuracy.

Fig. 4 (a) shows the simulated time taken across layers for the main-exit and early-exit paths. At the processor clock frequency of 3.2 MHz, processing an image takes 4.07 minutes for the main-exit path but only 1.27 minutes for the early-exit path, reducing the processing time and thus energy consumption by 68.8%. Computing the convolutional layer dominates the total processing time due to multiples of matrix multiplication in series. The first convolutional layer (CV1) costs longer time than the other following layers since the input size becomes smaller due to the pooling operation. Fig. 4 (b) shows allocated memory usage for different layer computation. The minimum remaining SRAM memory of 860 bytes in the processor chip can store 1,720 4-bit inference results, and additional results can be stored in the flash memory.

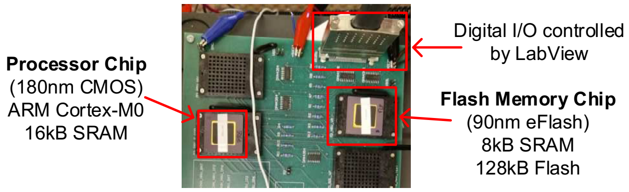


Fig. 6. Photo of the testing setup.

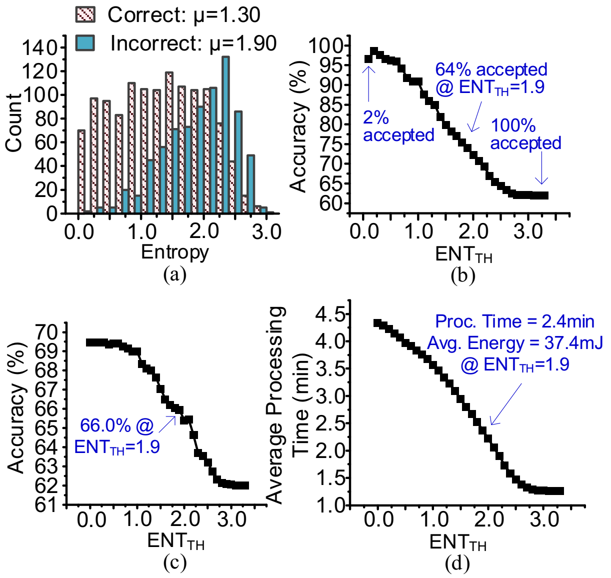


Fig. 7. Measured CNN for 2k images. (a) Entropy distribution. (b) Accuracy of only the early-exit path and acceptance ratio. (c) Accuracy of the early-exit path followed by the main-exit path if entropy > $\mathrm{ENT_{TH}}$. (d) Average processing time.

Fig. 5 (a) presents the bit assignment for the input and intermediate data, which is designed in 8 and 16 bits, respectively, to efficiently utilize the 32-bit SRAM. Fig. 5 (b) shows how the inference accuracy changes according to the number of decimal bits of the input image and layer coefficients. The decimal part is chosen to 5 bits to reduce the data from 16 to 8 bits while sacrificing the accuracy by only 1.68% and 2.44% for the main and early-exit paths, respectively.

V. Measurement Results

Fig. 6 shows the photo for the testing setup. The program code for the neural network operation and its coefficients are written to the SRAM of the processor chip or the flash memory by National Instruments PCIe-6535b digital I/O device. Inference and entropy results recorded by monitoring the bus by the same device. Its power consumption is measured by Keithley 2401/2450 sourcemeters.

Fig. 7 (a) presents the measured entropy distribution of the early- exit path using randomly selected 2000 images from the CIFAR-10/Test dataset. The average entropy of correct inferences is 1.30 while that of incorrect ones is 1.90. Fig. 7 (b) shows the inference accuracy when the results are accepted if entropy is lower than a threshold ($\mathrm{ENT_{TH}}$). Also, it shows how many inference results are accepted at three $\mathrm{ENT_{TH}}$ values. The number of inferenced images can be improved by running the main-exit path when the entropy is higher than $\mathrm{ENT_{TH}}$. However, it increases the total energy consumption. Figs. 7 (c) and (d) show the accuracy and the average processing time across $\mathrm{ENT_{TH}}$ for this approach. Note that the total time for re-inference, when entropy > $\mathrm{ENT_{TH}}$, is the processing time of 'early-exit path + main-exit path – shared part'. With $\mathrm{ENT_{TH}}$ of 1.9, it reduces processing time and thus energy consumption by 41.3% (63.7mJ → 37.4mJ) compared with the main-exit only method while lowering the accuracy from 69.5% to 66.0%.

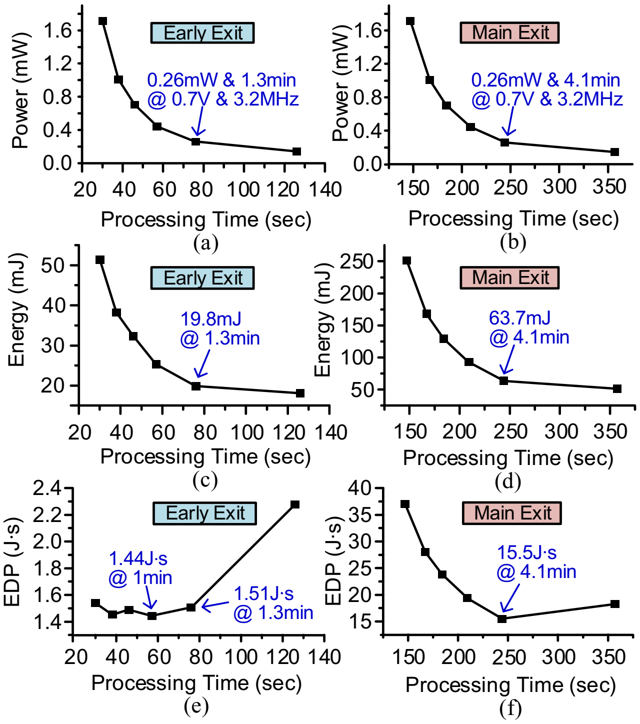


Fig. 8. Measured power and energy consumption across processing time. (a) Power consumption for the early-exit path. (b) Power consumption for the main-exit path. (c) Energy consumption for the early-exit path. (d) Energy consumption for the main-exit path. (e) EDP for the early-exit path. (f) EDP for the main-exit path.

Figs. 8 (a), (b), and (c) show the power consumption and total energy cost to perform an inference. The processing time is adjusted by supply voltage and internal oscillator configuration. Figs. (d) and (e) shows their energy-delay product (EDP). At 0.7 V supply voltage and 3.2 MHz clock frequency, it achieves near optimal EDP for both the main-exit and early-exit paths. Here, the early-exit path has the 3.2× shorter processing time and thus 3.2× lower energy consumption per inference than the main-exit path.

VI. Conclusion

This paper presents a neural network designed for a miniature edge device. It includes two exit options for dynamic available energy condition in energy-harvesting-powered system. The implemented neural network reduces processing time and thus energy consumption by 41.3% compared with the main-exit only method while sacrificing its accuracy from 69.5% to 66.0%.

REFERENCES

[1] Y. Lee, *et al*., "A Modular 1 mm^3 Die-Stacked Sensing Platform With Low Power I^{2}C Inter-Die Communication and Multi-Modal Energy Harvesting," *IEEE J. Solid-State Circuits,* vol. 48, no. 1, pp. 229-243, Jan. 2013.

[2] V. Iyer, A. Najafi, J. James, S. Fuller, and S. Gollakota, "Wireless steerable vision for live insects and insect-scale robots," *Science Robotics,* vol. 5, Issue 44, Jul. 2020 .

[3] Y. Chen, *et al*, "An Injectable 64 nW ECG Mixed-Signal SoC in 65 nm for Arrhythmia Monitoring," *IEEE JSSC,* vol. 50, no. 1, pp. 229-243, Jan. 2015.

[4] G. Kim, *et al.,* "A Millimeter-Scale Wireless Imaging System with Continuous Motion Detection and Energy Harvesting," *Proc. IEEE Symp. VLSI Technology,* Jun. 2014.

[5] A. H. Alavi, *et al*., "Self-charging and self-monitoring smart civil infrastructure systems: current practice and future trends," in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems*, Mar. 2019.

[6] M. H. Ghaed, *et al.,* "Circuits for a Cubic-Millimeter Energy-Autonomous Wireless Intraocular Pressure Monitor," *IEEE Trans. Circuits and Systems – I,* vol. 60, no. 12, pp. 3152 – 3161, Dec. 2013.

[7] The CIFAR-10 dataset https://www.cs.toronto.edu/~kriz/cifar.html.

[8] Q. Dong, *et al,* "A 1Mb Embedded NOR Flash Memory with 39µW Program Power for mm-Scale High-Temperature Sensor Nodes," *IEEE ISSCC Dig. Tech. Papers,* pp. 198-200, Feb. 2017.

[9] M. Diana, J. Chikama, M. Amagasaki, M. Iida, M. Kuga, "Characteristic Similarity Using Classical CNN Model," *34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, Jun. 2019.

[10] CYMBET Corporation, "Rechargeable Solid State Bare Die Batteries," EnerChip™ Bare Die CBC005 Datasheet, 2016.

[11] Texas Instruments, "Mixed Signal Microcontroller," MSP430G2x11 MSP430G2x01 Datasheet, 2013.

[12] I. Lee, *et al.,* "A 179-lux Energy-Autonomous Fully-Encapsulated 17-mm^3 Sensor Node with Initial Charge Delay Circuit for Battery Protection," *IEEE Symp. VLSI Technology Dig. Tech. Papers,* pp. 251-252*,* Jun. 2018.

[13] W. Jung, *et al.,* "A 3nW Fully Integrated Energy Harvester Based on Self-Oscillating Switched-Capacitor DC-DC Converter," *IEEE ISSCC Dig. Tech. Papers,* pp. 398-400, Feb. 2014.

[14] I. Lee, *et al.,* "A >78%-Efficient Light Harvester over 100-to100klux with Reconfigurable PV-Cell Network and MPPT Circuit," *IEEE ISSCC Dig. Tech. Papers,* pp. 370-372, Feb. 2016.

[15] Y. Wu, Z. Wang, Z. Jia, Y. Shi, and J. Hu, "Intermittent Inference with Nonuniformly Compressed Multi-Exit Neural Network for Energy Harvesting Powered Devices," ACM/*IEEE DAC,* Jul. 2020.

[16] V. Iyer, R. Nandakumar, A. Wang, S. B. Fuller, and S. Gollakota, "Living IoT: A Flying Wireless Platform on Live Insects," *25th Annual International Conference on Mobile Computing and Networking,* Aug. 2019.

[17] S. Thomas, R. R. Harrison, A. Leonardo, and M. S. Reynolds, "A Battery-Free Multichannel Digital Neural/EMG Telemetry System for Flying Insects," *IEEE Trans. Biomedical Circuits and Systems,* vol. 6, no. 5, pp. 424-436, Oct. 2012.

[18] I. Park *et al.,* "A 640 × 640 Fully Dynamic CMOS Image Sensor for Always-On Operation" *IEEE JSSC,* vol. 55, no. 4, pp. 898-907, Apr. 2020.

[19] Y. Kim *et al.*, "A 1/3-inch 1.12µm-Pitch 13Mpixel CMOS Image Sensor with a Low-power Readout Architecture," *IEEE ISCAS*, Oct. 2020.

[20] K. D. Choo *et al,* "Energy-Efficient Low-Noise CMOS Image Sensor with Capacitor Array-Assisted Charge-Injection SAR ADC for Motion-Triggered Low-Power IoT Applications", *IEEE ISSCC Dig. Tech. Papers,* pp. 96-98, Feb. 2019.

[21] W. Zhao, C. Park, I. Park, N. Sun, and Y. Chae, "An Always-On 4× Compressive VGA CMOS Imager with 51pJ/pixel and >32dB PSNR," *IEEE Symp. VLSI Circuits,* Jun. 2020.

[22] A. Marzuki, Z. A. A. Aziz, and A. B. Manaf, "A Review of CMOS Analog Circuits for Image Sensing Application", *IEEE International Conf. Imaging Systems and Techniques*, May 2011.

[23] I. F. Akyildiz, T. Melodia, and K. R. Chowdury, "Wireless multimedia sensor networks: A survey," *IEEE Wireless Comuniations*, Dec. 2007

[24] A. N. Belbachir, *Smart Cameras*, Vienna, Austria; Springer, 2010.

[25] I. Civek and S. U. Ay, "An ultra-low power energy harvesting and imaging (EHI) type CMOS APS imager with self-power capability," *IEEE Trans. Circuits and Systems – I,* vol. 62, Issue 9, pp. 2177 – 2186, Sep. 2015.

[26] L. Chuo, *et al.,* "A 915MHz Asymmetric Radio Using Q-Enhanced Amplifier for a Fully Integrated 3×3×3mm^3 Wireless Sensor Node with 20m Non-Line-of-Sight Communication," *IEEE ISSCC Dig. Tech. Papers,* pp. 132-134, Feb. 2017.

[27] H. An, *et al.,* "A 170µW Image Signal Processor Enabling Hierarchical Image Recognition for Intelligence at the Edge", *IEEE Symp. VLSI Circuits,* Jun. 2020.

[28] S. Jeong, Y, Kim, G. Kim, and D. Blaauw, "A Pressure Sensing System with ±0.75 mmHg (3σ) Inaccuracy for Battery-Powered Low Power IoT applications", *IEEE Symp. VLSI Circuits,* Jun. 2020.

[29] S. Oh, *et al.,* "A 2.5nJ Duty-Cycled Bridge-to-Digital Converter Integrated in a 13mm^3 Pressure-Sensing System," *IEEE ISSCC Dig. Tech. Papers,* pp. 328-330, Feb. 2018.

[30] I. Lee, E. Moon, Y. Kim, J. Phillips, and D. Blaauw, "A 10mm^3 Light-Dose Sensing IoT2 System with 35-to-339nW 10-to-300klx Light-Dose-to-Digital Converter," *IEEE Symp. VLSI Circuits,* Jun. 2019.

[31] T. Kang, *et al.,* "A 1.74.12 mm^3 Fully Integrated pH Sensor for Implantable Applications using Differential Sensing and Drift-Compensation," *IEEE Symp. VLSI Circuits,* Jun. 2019.

[32] M. Cho, *et al.,* "A 142nW Voice and Acoustic Activity Detection Chip for mm-Scale Sensor Nodes Using Time-Interleaved Mixer-Based Frequency Scanning," *IEEE ISSCC Dig. Tech. Papers,* pp. 278-280, Feb. 2019.

[33] M. Cho, *et al.,* "A 6×5×4mm^3 General Purpose Audio Sensor Node with a 4.7µW Audio Processing IC," *IEEE Symp. VLSI Circuits,* Jun. 2017.

[34] K. Choo, *et al,* "Energy-Efficient Motion-Triggered IoT CMOS Image Sensor With Capacitor Array-Assisted Charge-Injection SAR ADC," *IEEE JSSC,* vol. 54, no. 11, pp. 2921-2931, Nov. 2019.

[35] P. Pannuto, *et al.,* "MBus: An Ultra-Low Power Interconnect Bus for Next Generation Nanopower Systems," ACM/IEEE ISCA, Jun. 2015

[36] W. Jung, D. Sylvester, D. Blaauw, "A Rational-Conversion-Ratio Switched-Capacitor DC-DC Converter Using Negative-Output Feedbac," *IEEE ISSCC Dig. Tech. Papers,* pp. 218-220, Feb. 2016.

[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *IEEE/CVF Conf. Computer Vision and Pattern Recognition,* Jun. 2018.

[38] G. Gobieski, B. Lucia, and N. Beckmann, "Intelligence beyond the edge: Inference on intermittent embedded systems," *ACM Proc. Twenty-Fourth International Conf. Architectural Support for Programming Languages and Operating Systems,* pp. 199-213, Apr. 2019.

[39] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Branchynet: Fast inference via early exiting from deep neural networks," *IEEE ICPR,* Dec. 2016.

[40] G. Huang, D. Chen, T. Li, F. Wu, L. V. D. Maaten, and K. Q. Weinberger, "Multi-scale dense convolutional networks for efficient prediction," arXiv preprint arXiv:1703.09844, vol. 2, 2017.