An Ultralow-Power H.264/AVC Intra-Frame Image Compression Accelerator for Intelligent Event-Driven IoT Imaging Systems

Qirui Zhang, *Member, IEEE*, Hyochan An, Andrea Bejarano-Carbo, Hun-Seok Kim, *Senior Member, IEEE*, David Blaauw, *Fellow, IEEE*, and Dennis Sylvester

***Abstract*—This letter presents an ultralow-power (ULP) H.264/AVC intra-frame image compression accelerator tailored for intelligent event-driven ULP IoT imaging systems. The H.264/AVC intra-frame codec is customized to enable compression of arbitrary nonrectangular change-detected regions. To optimize energy and latency from image memory accesses, novel algorithm-hardware co-designs are proposed for intra-frame predictions, reducing overhead for neighbor macroblock (McB) accesses by 2.6× at a negligible quality loss. With split control for major processing phases, latency is optimized by exploiting data dependency and pipelining. Area and leakage of major computation units are reduced through data path micro-architecture reconfiguration. Fabricated in 40 nm, it occupies a mere 0.32 mm^2 area with 4-kB SRAM. At 0.6 V and 153 kHz, it consumes only 1.21 μW, with 30.9 pJ/pixel compression energy efficiency that rivals state-of-the-art designs. For an event-driven IoT imaging system, the combination of the proposed accelerator and change detection brings 133× reduction to the overall energy for regressing an image of change-detected region of interest.**

***Index Terms*—Algorithm-hardware co-design, event-driven imaging system, H.264/AVC intra-frame coding, hardware accelerator, ultralow-power (ULP).**

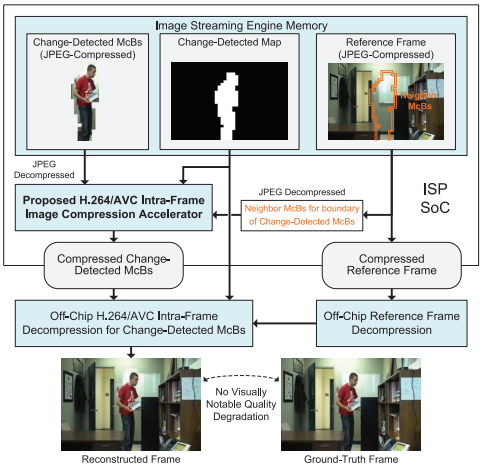

Fig. 1. Customized H.264/AVC intra-frame codec to support change detection.

I. Introduction

Pervasively in our daily lives, standards, such as H.264/AVC [1] and H.265 [2], enable high-speed live streaming and high-capacity storage of images and videos and have given rise to the multimedia era. The H.264/AVC intra-frame coding stands as one of the most effective image compression algorithms, which provides higher iso-quality compression ratio than its predecessors like H.263 [3], and lower-computation complexity than its successors like H.265, making it an attractive candidate for enabling compression in IoT imaging systems. With most previous H.264/AVC intra-frame compression accelerators [4], [5], [6] focusing on the optimization of performance, few designs have been tailored for IoT imaging systems, which enable critical applications like *in vivo* medical monitoring and security surveillance [7]. In such systems, image compression is indispensable for increasing storage capacity and reducing wireless data transmission cost [7], [8], [9], [10].

In this letter, we present an H.264/AVC intra-frame image compression accelerator designed as a block in an image signal processing (ISP) system-on-chip (SoC) [8], [9] serving as the "brain" of an intelligent event-driven ultralow-power (ULP) IoT imaging system [7]. The accelerator is tailored to support compression of arbitrary nonrectangular change-detected regions with ultralow-power and high-energy efficiency.

Manuscript received 4 August 2023; revised 30 November 2023; accepted 17 December 2023. Date of publication 20 December 2023; date of current version 16 January 2024. This work was supported by Sony Semiconductor Solutions Corporation/Sony Electronics Inc. This article was approved by Associate Editor Priyanka Raina. *(Corresponding author: Qirui Zhang.)*

Qirui Zhang, Andrea Bejarano-Carbo, Hun-Seok Kim, David Blaauw, and Dennis Sylvester are with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: qiruizh@umich.edu).

Hyochan An is with Apple, Cupertino, CA 95014 USA.

Digital Object Identifier 10.1109/LSSC.2023.3344699

II. Algorithm and Hardware Design

A. Algorithm-Hardware Co-Designs

To support compression of arbitrary nonrectangular change-detected regions, the H.264/AVC intra-frame codec is customized, as is shown in Fig. 1. The codec assumes static background of images for its typical use cases. For a sequence of input images, the ISP SoC stores the first as the reference frame, and stores only change-detected regions of subsequent frames, all in JPEG-compressed format. To faithfully predict values for boundary macroblocks (McBs) of a change-detected region, ideally their neighbor McBs from the same frame are required. However, in our proposed approach, the neighbors have been discarded when images are being streamed in, as they are nonchange-detected. Nevertheless, given the nature of change detection, nonchange-detected McBs in subsequent frames are supposed to be numerically close to corresponding ones in the reference frame. As such, in our customized codec, the reference frame is used to supply neighbor McBs for the boundary of change-detected regions to provide predictions that are as faithful as using those from the same frame. In cases where the background changes significantly, the ISP SoC can label the current frame as the new reference frame and the proposed accelerator compresses the full frame without using the previous reference.

Fig. 1 also shows a qualitative experimental result for the customized codec. Except that nonchange-detected pixels in the reconstructed frame can be slightly different from the ones in ground-truth, the change-detected region in the reconstructed frame does not exhibit visually notable image quality degradation when compared to the groundtruth.

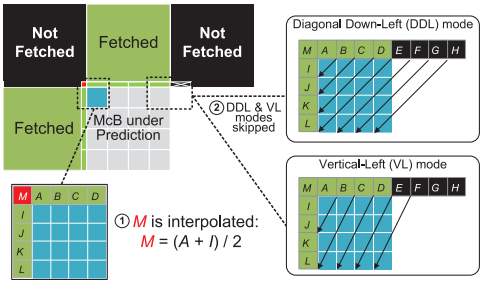

Fig. 2. Algorithm-hardware co-designs for intra-frame predictions.

In the ISP SoC, access to any pixel in a reference McB requires JPEG decompression for the entire McB. Shown in Fig. 2, intra-frame prediction in the baseline algorithm [11] needs to access four neighbor McBs, where only one pixel is used for the upper-left one and only four pixels are used for the upper-right one. Thus, accessing the upper-left and upper-right neighbors is highly cost-inefficient and induces substantial latency and energy overhead. To reduce that cost, algorithm-hardware co-designs are proposed, as is shown in Fig. 2. In the co-design, the upper-left and upper-right neighbors are not fetched. For the upper-left neighbor, adjacent pixels from upper and left neighbors are averaged to interpolate the missing pixel "*M*," which is necessary for three Luma prediction modes. For the upper-right neighbor, prediction modes involving its pixels are simply skipped for the upper-right 4×4 Luma block. In an experiment where change-detected regions are compressed for a sequence of 2000 frames, the co-designs bring an average 2.6× reduction for neighbor McB accesses, from 100 McBs per frame to only 38 McBs per frame, where the average number of change-detected McBs per frame is 155 (observed from the test dataset). That brings down latency and energy cost from the SoC's JPEG engine and bus transactions also by 2.6×. Note that, interpolating *M* makes the compression no longer compatible with the standard H.264/AVC intra-frame decompression. Hence, we also incorporated corresponding changes into the MATLAB-based off-chip decompression shown in Fig. 1.

B. Accelerator Architecture

Fig. 3 shows the architecture of the proposed accelerator. The accelerator is designed as an SoC block communicating with other blocks through ARM's AMBA AHB interconnect, and is capable of compressing colored images up to VGA (640 × 480) resolution. To achieve a balanced design with moderate performance and low hardware cost for ULP scenarios, 4-way parallelism is used, where four pixels from one row of a 4×4 block are processed in parallel.

Fig. 4 shows the scheduling of the accelerator. The major processing phases (intra-frame prediction, transform and quantization, and entropy coding) each have their own controller, enabling exploitation of data dependency and pipelining for optimized latency. Mode selection runs in parallel with the prediction data path and outputs the best prediction mode immediately after the prediction ends. To avoid buffering predicted pixels for all nine modes, the prediction data path runs the best prediction mode again after mode selection, where forward transform and quantization follow immediately as pipeline stages. After forward quantization, entropy coding starts immediately, with inverse quantization, inverse transform and reconstruction being computed in parallel. Immediately after reconstruction, the next block starts without waiting for the entropy coding of the previous block to end, as the tail of entropy coding latency can be hidden under the prediction latency of the next. The two 8×8 Chroma blocks are compressed in a similar manner. Without the optimizations, it takes > 130 cycles to compress a Luma block. The optimizations bring

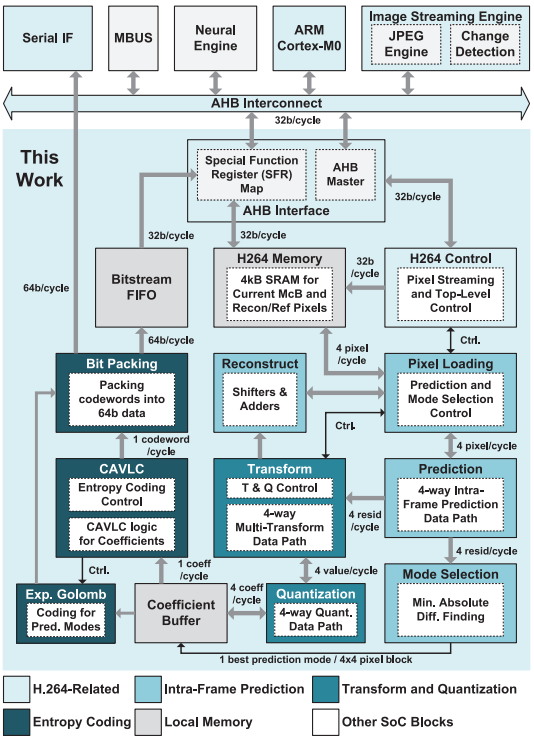

Fig. 3. Architecture of the proposed accelerator.

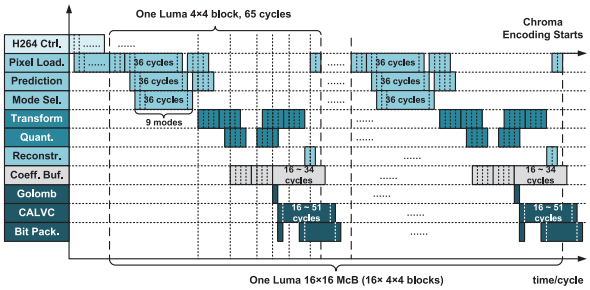

Fig. 4. Scheduling of the proposed accelerator.

that latency to only 65 cycles, a > 2× reduction. On average, it takes 1500 cycles for the accelerator to compress one colored McB.

C. Cost-Efficient Data Path Micro-Architectures

As ULP imaging scenarios are typically not critically sensitive to latency but to energy and cost, the proposed accelerator is designed to have a moderate performance but low energy and cost. In that regard, leakage power and area are minimized mainly through exploiting reconfigurable data path micro-architectures for major computation units (predictions, transforms, and quantizations), where similar computation kernels reuse the same arithmetic units, wiring, and control without incurring notable overhead, a methodology similar to the processing element design in [12]. The first practice is the reconfigurable intra-frame prediction data path design shown in Fig. 5. The H.264/AVC intra-frame prediction has nine modes for Luma and four modes for Chroma, and for a single mode, each pixel in a block is further predicted using different formulas, which incur hundreds of different computation patterns. However, many of the patterns share the same type of arithmetic units, such as adders and shifters, thus a reconfigurable data path can be designed. By selecting input pixels from the buffer for neighbor pixels and configuring

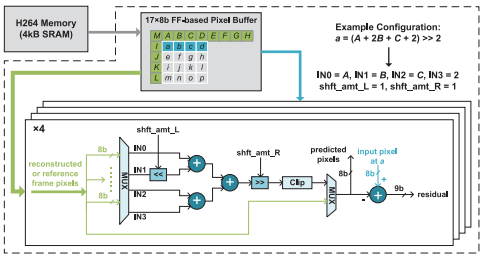

Fig. 5. Reconfigurable 4-way intra-frame prediction data path.

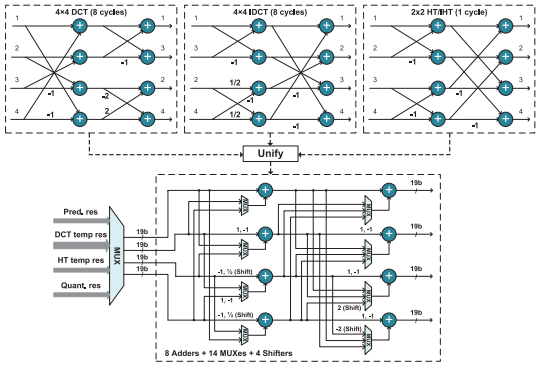

Fig. 6. Reconfigurable 4-way multitransform data path.

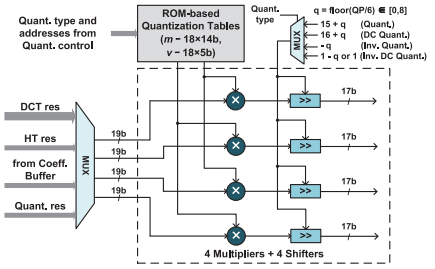

Fig. 7. Reconfigurable 4-way quantization data path.

parameters, such as the shifting amount, the prediction data path can support all the intra-frame prediction patterns.

Further exploration of the transforms leads to the design of a multitransform data path that leverages threefold hardware reusing, as shown in Fig. 6. The first fold is the intrinsic reuse of intermediate results for discrete cosine transform (DCT) and Hadamard transform (HT), which is enabled by the butterfly structure that is well-known for being used in fast Fourier transform. For DCT, the hardware cost for multiplying one row (column) of the input matrix with the right (left) 4×4 twiddle factor matrix is reduced from 12 additions to 8 additions. The second fold is intratransform reusing between the left twiddle factor matrix multiplication and the right one, where the two can share the same data path at the overhead of transposing the intermediate results after the right multiplication (to be fed to the left multiplication). In our design, that is simply achieved through reading intermediate results from a buffer in the transposed pattern, instead of a true transpose. The third fold is intertransform reusing, where DCT, inverse DCT, HT, and inverse HT have similar butterfly structures with the same number of arithmetic units (eight adders). The three butterfly structures are thus unified into a multitransform data path which can be reused to compute all four transforms by simply configuring the multiplexers.

As Fig. 7 shows, the quantization data path is also designed with the aforementioned methodology that exploits hardware reusing and

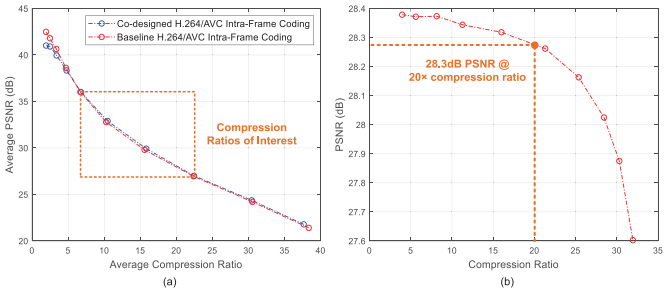

Fig. 8. Experimental results for decompressed image quality. (a) Comparison between the co-designed and baseline intra-frame codec. (b) Measurement results of a change detection region.

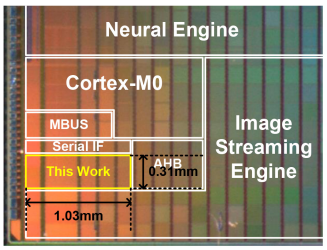

Fig. 9. Die photograph showing the proposed accelerator and other on-chip blocks.

reconfiguration. All different types of forward/inverse quantization shown in Fig. 7 (top right) involves first multiplying the inputs (transformed residuals or quantized coefficients) with constant factors defined by the H.264 intra-frame standard, then shifting the results by certain amount. Based on that observation, the quantization data path reuses four pairs of multipliers and shifters to support all four types of quantization. The constant factors are supplied to the multipliers through configuring addresses to the ROM-based quantization tables. The shifting amount is selected based on the quantization type and quantization parameter (QP).

III. Experimental Results

Fig. 8(a) shows decompressed image quality comparison between the co-designed and baseline intra-frame codec for full VGA frames across various scenes. For compression ratios $\leq 5\times$, the co-design (essentially a simplified variant of the H.264 intra-frame codec) exhibits lower image quality due to the proposed changes. For the $10\times \sim 20\times$ compression ratio range, no image quality degradation is observed for the co-design. Remarkably, image quality for the co-design can even surpass that of the baseline, suggesting the use of interpolated pixel M may offer more accurate predictions.

Fig. 8(b) shows measurement results for the decompressed image quality of change-detected region (an intruder) from a realistic image. At $20\times$ compression ratio for the change-detected region, the codec exhibits 28.3-dB PSNR. In the tested image, only 208 out of 1200 McBs are change-detected. When considering the compressed change-detected region against a full frame, $>115\times$ system-level compression ratio can be achieved, a significant saving for resource-constrained IoT imaging applications.

The proposed accelerator is fabricated in 40-nm CMOS technology. A custom ultralow leakage design [13], [14] is used for the 4-kB SRAM. Fig. 9 shows the die photograph for the proposed accelerator and other SoC blocks. The accelerator occupies a mere 0.32 mm^2 area with 1.03×0.31 mm footprint. With the SoC occupying 27 mm^2 area, the proposed accelerator provides high-efficiency image compression with only 1.2% area in the full SoC [8], [9].

TABLE I
COMPARISON WITH STATE-OF-THE-ART

	[1]TCSVT 2009 [4]	[1]TVLSI 2011 [5]	VLSI 2012 [6]	This work
Technology	130nm	130nm	65nm	40nm
Area	0.72mm^2	3.17mm^2	2.07mm^2	0.32mm^2
SRAM	1.8kB	6.4kB	27.1kB	4kB
Resolution	1080p	1080p	1080p ∼ 4320p	640×480
Prediction Modes	4×4/16×16	4×4/16×16 /Plane	8×8/16×16	4×4
Entropy Coding	CAVLC	CABAC	CABAC	CAVLC
Voltage	1.2V	1.2V	0.8V ∼ 1.2V	0.6V ∼ [1]1.1V
Frequency	140MHz	114MHz	9MHz ∼ 260MHz	153kHz ∼ [1]100MHz
Performance	30fps	30fps	30fps 1080p @ 9MHz 60fps 4320p @ 260MHz	102McB/s @ 153kHz [1]56fps @ 100MHz
Power	Not reported	23.56mW	2.0mW ∼ 106.7mW	1.21μW @ 0.6V, 153kHz
Energy/Pixel	Not reported	378.7pJ	24.5pJ @ 0.8V, 33MHz 53.6pJ @ 1.2V, 260MHz	30.9pJ @ 0.6V, 153kHz
[2]Scaled Power	N/A	31.5μW @ 0.6V, 2.5MHz	335.53μW @ 0.6V, 2.5MHz	14.53μW @ 0.6V, 2.5MHz
[2]Scaled Energy/Pixel	N/A	23.1pJ @ 0.6V, 2.5MHz	19.4pJ @ 0.6V, 2.5MHz	22.7pJ @ 0.6V, 2.5MHz

[1]Simulation. [2]Technologies in [5] and [6] are scaled to 40nm CMOS using data from [15].

TABLE II
IMAGING SYSTEM GAINS FROM USING THE PROPOSED ACCELERATOR

Method	Frame Size	Proc. Energy	Flash Energy	Egress Energy	Egress Time	Capacity (# Frames)
No Comp.	3.7Mb	0	407μJ	414mJ	2 hours	0
JPEG	335kb	11.9μJ	36.9μJ	37.7mJ	11.6 min	6
H.264	158.2kb	28.3μJ	17.4μJ	17.8mJ	5.5 min	12
CD & H.264	27.4kb	14.8μJ	3μJ	3.1mJ	57s	74

Table I shows the summary of chip measurement results and comparison with state-of-the-art. Measured at 0.6 V and 153 kHz, the proposed accelerator consumes only 1.21-μW power. With an average of 1500 cycles to compress a colored McB (384 pixels), it achieves 30.9-pJ/pixel energy efficiency at 0.6 V and 153 kHz. When implemented and simulated at 1.1 V, the accelerator is able to operate at 100 MHz, delivering 56 fps peak performance for VGA images. For fair comparison, we further scale the technologies in [5] and [6] to 40 nm, and report the scaled power and energy efficiency at 0.6 V and 2.5 MHz (fastest clock frequency for the SoC at 0.6 V). Though the primary goal of the proposed design is to minimize power, not necessarily performance or efficiency, its energy efficiency still rivals state-of-the-art designs [5], [6] mainly due to the lower complexity of our simplified intra-frame prediction. It is worth noting that the proposed design can exhibit much lower power than the other designs, making it more desirable for the ULP IoT imaging scenarios targeted in this letter.

Table II shows the gains for an intelligent millimeter-scale IoT imaging system [7] when using the proposed accelerator. The size of flash memory in this millimeter-scale system is only 2 Mb. Without compression techniques, the system cannot even hold a single colored VGA frame, and egressing the raw frame through an RF transmitter consumes unacceptably high energy and long latency that soon kills the battery. With H.264/AVC applied to full frames, 23× compression ratio is achieved, which boosts system storage capacity to 12 VGA frames, reduces the overall system energy by 23× and the egress latency by 22×. Though H.264/AVC processing energy is higher than JPEG, the overall system energy when using H.264/AVC is $> 2\times$ lower, due to higher compression ratio of H.264/AVC and the fact that RF wireless egressing dominates the overall system energy. With the synergy of change detection and H.264/AVC, the compression ratio can be boosted to 135× and the system storage capacity is raised to remarkably 74 VGA frames. The overall system energy is drastically reduced by 133×, with the egress latency reduced by 126× to only 57 s, ensuring timely reports of potential intruders in security surveillance.

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003, doi: 10.1109/TCSVT.2003.815165.

[2] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—Including high efficiency video coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1669–1684, Dec. 2012, doi: 10.1109/TCSVT.2012.2221192.

[3] A. Joch, F. Kossentini, H. Schwarz, T. Wiegand, and G. J. Sullivan, "Performance comparison of video coding standards using Lagrangian coder control," in *Proc. Int. Conf. Image Process.*, Rochester, NY, USA, 2002, pp. II.501–II.504, doi: 10.1109/ICIP.2002.1039997.

[4] Y.-K. Lin, C.-W. Ku, D.-W. Li, and T.-S. Chang, "A 140-MHz 94 K gates HD1080p 30-frames/s intra-only profile H.264 encoder," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 432–436, Mar. 2009, doi: 10.1109/TCSVT.2009.2013511.

[5] H.-C. Kuo, L.-C. Wu, H.-T. Huang, S.-T. Hsu, and Y.-L. Lin, "A low-power high-performance H.264/AVC intra-frame encoder for 1080pHD video," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 6, pp. 925–938, Jun. 2011, doi: 10.1109/TVLSI.2010.2045402.

[6] D. Zhou, G. He, W. Fei, Z. Chen, J. Zhou, and S. Goto, "A 4320p 60fps H.264/AVC intra-frame encoder chip with 1.41Gbins/s CABAC," in *Proc. Symp. VLSI Circuits*, Honolulu, HI, USA, 2012, pp. 154–155, doi: 10.1109/VLSIC.2012.6243836.

[7] A. Bejarano-Carbo et al., "Millimeter-scale ultra-low-power imaging system for intelligent edge monitoring," in *Proc. tinyML Res. Symp.*, 2022, pp. 1–7, doi: arxiv.org/abs/2203.04496.

[8] H. An et al., "A 170μW image signal processor enabling hierarchical image recognition for intelligence at the edge," in *Proc. IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, 2020, pp. 1–2, doi: 10.1109/VLSICircuits18222.2020.9162810.

[9] H. An et al., "An ultra-low-power image signal processor for hierarchical image recognition with deep neural networks," *IEEE J. Solid-State Circuits*, vol. 56, no. 4, pp. 1071–1081, Apr. 2021, doi: 10.1109/JSSC.2020.3041858.

[10] Z. Fan et al., "Audio and image cross-modal intelligence via a 10TOPS/W 22nm SoC with back-propagation and dynamic power gating," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Honolulu, HI, USA, 2022, pp. 18–19, doi: 10.1109/VLSITechnologyandCir46769.2022.9830226.

[11] A. A. Muhit. "MATLAB central file exchange." 2023. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/39927-h-264-baseline-codec

[12] Q. Zhang et al., "A 22nm 3.5TOPS/W flexible micro-robotic vision SoC with 2MB eMRAM for fully-on-chip intelligence," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Honolulu, HI, USA, 2022, pp. 72–73, doi: 10.1109/VLSITechnologyandCir46769.2022.9830340.

[13] J. Wang, H. An, Q. Zhang, H. S. Kim, D. Blaauw, and D. Sylvester, "1.03pW/b ultra-low leakage voltage-stacked SRAM for intelligent edge processors," in *Proc. IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, 2020, pp. 1–2, doi: 10.1109/VLSICircuits18222.2020.9162843.

[14] J. Wang, H. An, Q. Zhang, H. S. Kim, D. Blaauw, and D. Sylvester, "A 40-nm ultra-low leakage voltage-stacked SRAM for intelligent IoT sensors," *IEEE Solid-State Circuits Lett.*, vol. 4, pp. 14–17, 2021, doi: 10.1109/LSSC.2020.3043461.

[15] A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance from 180nm to 7nm," *Integration*, vol. 58, pp. 74–81, Jun. 2017, doi: 10.1016/j.vlsi.2017.02.002.