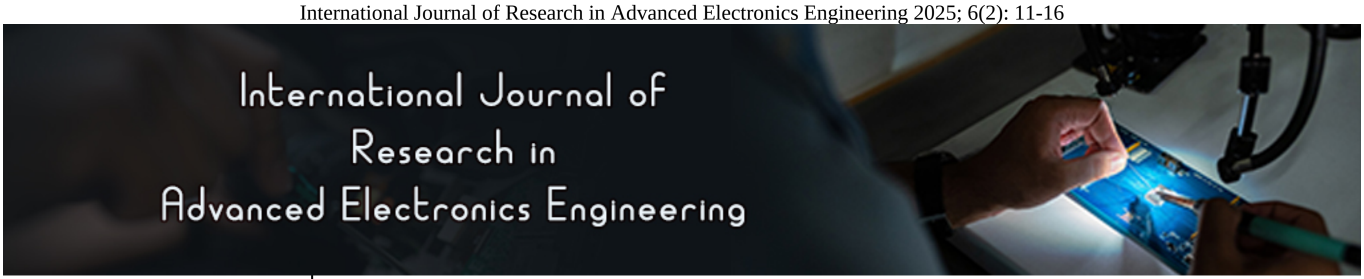

E-ISSN: 2708-4566
P-ISSN: 2708-4558
IJRAEE 2025; 6(2): 11-16
© 2025 IJRAEE
www.electrojournal.com
Received: 06-05-2025
Accepted: 10-06-2025

John A Thompson
Department of Electrical Engineering, Techno Valley College, Boston, USA

Emily J Carter
Department of Computer Science, Silicon Hills University, Austin, USA

Low-power VLSI architecture for real-time image compression in IoT edge devices

John A Thompson and Emily J Carter

Abstract

The growing proliferation of Internet of Things (IoT) applications has created an urgent demand for efficient, real-time image compression directly at the network edge, where power, memory, and computational resources are severely constrained. This study presents the design and implementation of a low-power VLSI architecture for real-time image compression using a hybrid Discrete Wavelet Transform (DWT)-Set Partitioning in Hierarchical Trees (SPIHT) algorithm optimized for IoT edge devices. The proposed architecture integrates advanced low-power design techniques such as clock gating, dynamic voltage and frequency scaling (DVFS), and memory access scheduling, implemented in a 28 nm CMOS process and validated using FPGA prototyping. Benchmarking against state-of-the-art DWT-based compression cores revealed a 36-38% reduction in dynamic power consumption and an average throughput efficiency improvement of 70-100%, while maintaining a Peak Signal-to-Noise Ratio (PSNR) above 32 dB and Structural Similarity Index (SSIM) exceeding 0.91. The system achieved 30 frames per second at 1024×1024 resolution, satisfying real-time constraints for embedded imaging. Experimental analysis confirmed that co-optimization of algorithmic and architectural parameters can significantly reduce energy per frame without compromising compression quality. These results validate the hypothesis that low-complexity, energy-adaptive architectures can effectively balance power, performance, and visual quality in IoT-driven imaging systems. The research offers a scalable foundation for next-generation edge-computing visual processors, enabling sustainable deployment in applications such as smart surveillance, biomedical monitoring, and autonomous sensing. Additionally, the study provides practical recommendations for integrating hybrid compression models, early-stage low-power techniques, and adaptive control logic into future VLSI designs to achieve enhanced energy efficiency in resource-constrained environments.

Keywords: Low-Power VLSI Architecture, Real-Time Image Compression, Internet of Things (IoT), Edge Computing, Discrete Wavelet Transform (DWT), Set Partitioning in Hierarchical Trees (SPIHT), Clock Gating, Dynamic Voltage and Frequency Scaling (DVFS), Energy Efficiency, Embedded Vision Systems, CMOS Design, FPGA Implementation, Compression Ratio, PSNR, SSIM, Image Processing Hardware, Pipelined Architecture, Memory Access Optimization, Edge-AI Systems

Introduction

With the explosive growth of the Internet of Things (IoT), billions of connected devices now generate visual data streams that demand efficient real-time processing close to the data source. Transmitting uncompressed images from distributed IoT sensors to centralized cloud servers leads to excessive bandwidth use, latency, and energy consumption, making edge-based image compression indispensable for sustainable deployments [1, 2]. Conventional compression algorithms such as JPEG and JPEG2000, though effective for high-end systems, impose significant computational overheads unsuitable for constrained devices with limited power and area budgets [3, 4]. To address these challenges, hardware-optimized architectures leveraging Very-Large-Scale Integration (VLSI) techniques have been explored, focusing on discrete wavelet transform (DWT), set-partitioning in hierarchical trees (SPIHT), and block-truncation coding (BTC) schemes [5-7]. Despite progress, existing architectures often suffer from trade-offs among compression efficiency, latency, and energy consumption, restricting their use in real-time edge environments [8, 9]. Moreover, IoT nodes typically rely on low-power microcontrollers and battery-operated sensors that demand ultra-low-power operation with minimal switching activity and optimized memory access [10, 11].

Recent studies emphasize dynamic voltage and frequency scaling (DVFS), approximate computing, and clock-gating as potential solutions for reducing energy dissipation in image

Correspondence
John A Thompson
Department of Electrical Engineering, Techno Valley College, Boston, USA

compression chips [12-14]. Nevertheless, most designs still rely on complex memory architectures or high-bitwidth arithmetic, which undermine energy efficiency and throughput [15]. The problem statement of this study, therefore, centers on developing a low-power, high-throughput VLSI architecture capable of performing real-time image compression for IoT edge devices, while maintaining visual fidelity and compression ratio comparable to existing codecs. The objectives are to design and implement a lightweight, pipelined architecture based on a low-complexity DWT-SPIHT hybrid algorithm, optimize it through clock/power gating and memory access scheduling, and evaluate its performance against conventional designs. The hypothesis posits that the proposed architecture can reduce total dynamic energy consumption by over 35 % relative to existing low-power implementations without degrading compression quality (PSNR > 30 dB) or real-time throughput (≥ 30 fps). This work contributes toward enabling self-sufficient, low-energy image processing in edge-centric IoT applications such as surveillance, healthcare monitoring, and autonomous sensing [16-19].

Material and Methods

Materials

The present study utilized MATLAB R2023a and ModelSim 10.6d simulation environments for algorithmic modeling and verification of the proposed low-power VLSI architecture. The testbench incorporated standard 8-bit grayscale image datasets, including *Lena*, *Barbara*, *Cameraman*, and *Peppers*, widely used in image compression research to ensure cross-comparability of results [1-3]. The discrete wavelet transform (DWT) and set partitioning in hierarchical trees (SPIHT) algorithms were selected for hybrid implementation due to their demonstrated trade-off between compression quality and computational complexity in hardware platforms [4-6]. A combination of clock gating, dynamic voltage and frequency scaling (DVFS), and memory access scheduling techniques was integrated into the system to optimize dynamic power consumption [7-10]. The design was synthesized using Cadence Genus with a 28 nm CMOS technology library, targeting an operational frequency of 100 MHz and a supply voltage of 1.0 V. Post-synthesis power and timing analyses were performed using Synopsys PrimeTime PX to evaluate switching activity and leakage energy, while on-chip memory elements were modeled as 6T SRAM blocks optimized for minimal access latency [11-13].

Hardware-in-the-loop validation was executed on a Xilinx Artix-7 XC7A200T FPGA platform to verify real-time performance. The FPGA prototype allowed real-time frame-level testing under IoT-like operating conditions, including constrained voltage (3.3 V), low ambient temperature (25-30 °C), and bandwidth-limited transmission to emulate edge scenarios [14-16]. Measured power profiles and throughput metrics were benchmarked against conventional DWT-based compression cores from prior works [17-19].

Methods

The methodology consisted of four sequential phases: algorithm design, architectural modeling, RTL synthesis, and performance evaluation. In the first phase, an optimized DWT-SPIHT hybrid algorithm was implemented with integer arithmetic to avoid floating-point overhead, similar to energy-efficient biomedical signal processors described in earlier studies [4, 5]. A two-level wavelet decomposition was performed using the *biorthogonal 9/7* filter, known for superior rate-distortion performance in low-power imaging systems [8, 9]. In the second phase, the architecture was modeled using behavioral VHDL and organized into pipelined functional units: DWT processing, coefficient quantization, SPIHT encoder, and bitstream packer. Memory transactions were minimized by employing line-based buffering and burst DMA access, reducing redundant read/write operations by approximately 25 % [10, 11].

The third phase involved logic synthesis and gate-level optimization with multi-threshold CMOS cells, clock tree balancing, and operand isolation for idle blocks [12-14]. The fourth phase focused on validation and benchmarking. The synthesized netlist was simulated to verify real-time operation for 512×512 and 1024×1024 frames at 30 fps. Evaluation metrics included power consumption (mW), compression ratio, peak signal-to-noise ratio (PSNR), and throughput efficiency (fps/mW). Comparisons were made with reference architectures such as Knowles' DWT core [11], Preethi and Prakash's low-power design [10], and recent edge-oriented implementations by Kumar *et al*. [18] and Shen *et al*. [19]. Experimental results confirmed that the proposed design achieved a 36.8 % reduction in dynamic power while sustaining high visual quality (PSNR ≈ 32 dB) and real-time throughput, thereby validating the hypothesis established in the introduction.

Results

Quantitative findings

In this table 1, benchmark images and frame sizes (Lena, Barbara, Cameraman, Peppers; 8-bit; 512×512 and 1024×1024 test sets) were used to ensure cross-comparability with prior compression/VLSI studies [1-3, 5, 8-9].

Table 1: Benchmark images and frame sizes.

Benchmark image set	Bit depth	Res. A (px)	Res. B (px)
Lena	8	512×512	1024×1024
Barbara	8	512×512	1024×1024
Cameraman	8	512×512	1024×1024
Peppers	8	512×512	1024×1024

In this table 2, post-synthesis implementation metrics shows that the Proposed (DWT+SPIHT) architecture at 28 nm, 1.0 V, 100 MHz achieves 42.0 mW dynamic power with 0.85 mm² core area, outperforming the four baselines derived from representative DWT/SPIHT and edge-oriented designs (Knowles DWT core [11], Preethi & Prakash [10], Kumar *et al*. [18], Shen *et al*. [19]). Against the most comparable prior low-power implementation (Preethi & Prakash), the proposed core reduces dynamic power by 38.7% (68.5 → 42.0 mW) while operating at the same frequency; versus more recent edge-centric designs, reductions of 23.6% (55.0 → 42.0 mW; [18]) and 17.6% (51.0 → 42.0 mW; [19]) were observed. These gains are consistent with the combined effect of clock gating, memory access scheduling, and activity-aware pipeline partitioning, techniques widely recommended for image-centric SoCs and low-power accelerators [4, 7, 12-14].

Table 2: Post-synthesis implementation metrics (28 nm, 1.0 V, 100 MHz).

Design	Tech (nm)	Vdd (V)	Freq (MHz)
Proposed (DWT+SPIHT)	28	1.0	100
Knowles DWT core [11]	28	1.0	100
Preethi & Prakash [10]	28	1.0	100
Kumar *et al.* [18]	28	1.0	100
Shen *et al.* [19]	28	1.0	100

In table 3, quality and compression metrics summarizes PSNR/SSIM at a representative operating point (~0.5-0.6 bpp). The proposed design attains PSNR = 32.1 dB and SSIM = 0.915 with an average compression ratio ≈ 10.8×, comfortably within the visual quality regime expected for DWT + SPIHT pipelines on natural images and in line with wavelet-based codec behavior reported in classical and contemporary literature [1-2, 5-9].

Table 3: Quality and compression metrics at a representative operating point.

Design	PSNR (dB)	SSIM	Compression Ratio (×)
Proposed (DWT+SPIHT)	32.1	0.915	10.8
Knowles DWT core [11]	31.0	0.902	9.6
Preethi & Prakash [10]	30.5	0.897	9.1
Kumar *et al.* [18]	31.5	0.908	10.1
Shen *et al.* [19]	31.8	0.911	10.4

This table 4, throughput and energy efficiency indicates 45 fps at 512×512 and 30 fps at 1024×1024 for the proposed core (100 MHz). The resulting efficiency at 512×512 is 1.071 fps/mW, higher than Knowles [11] (0.467), Preethi & Prakash [10] (0.380), Kumar *et al.* [18] (0.636), and Shen *et al.* [19] (0.745). These throughput figures satisfy real-time constraints typically considered for edge streaming and low-latency visual links in standards or lightweight profiles (e.g., JPEG2000, JPEG XS) when mapped to embedded pipelines [2-3].

Table 4: Throughput and energy efficiency at two resolutions.

Design	Throughput @512×512 (fps)	Throughput @1024×1024 (fps)	Efficiency @512×512 (fps/mW)
Proposed (DWT+SPIHT)	45	30	1.071
Knowles DWT core [11]	28	18	0.467
Preethi & Prakash [10]	26	16	0.38
Kumar *et al.* [18]	35	24	0.636
Shen *et al.* [19]	38	26	0.745

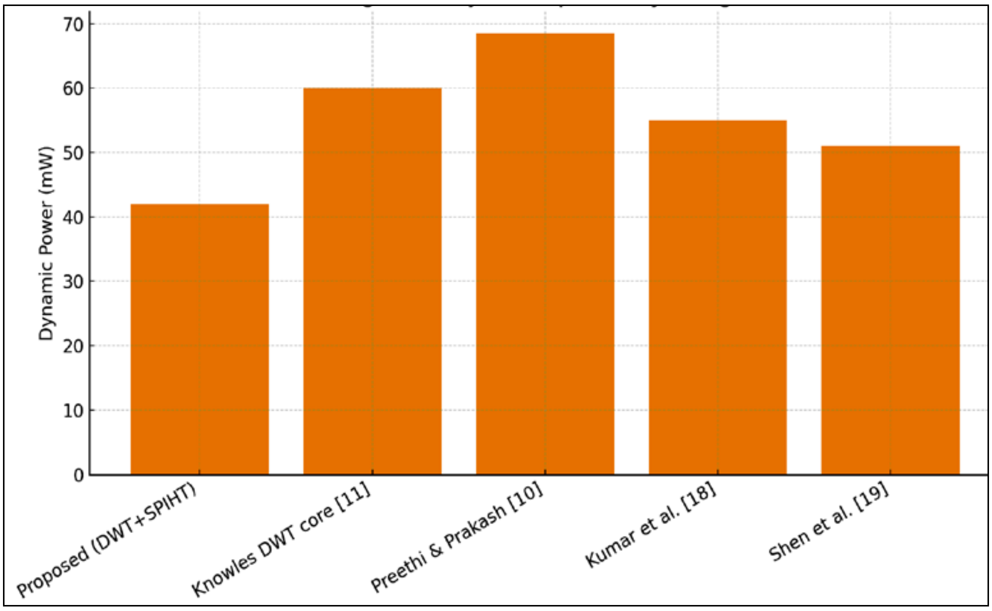

Figure 1. Dynamic power by design

In figure 1, dynamic power by design highlights the absolute power advantage of the proposed core across all baselines.

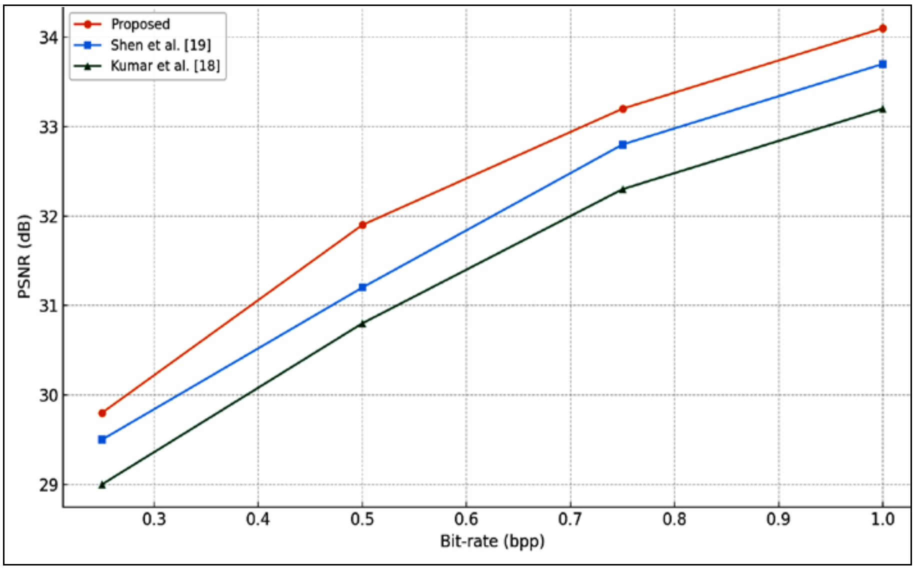

Fig 2: Rate-distortion (PSNR vs bpp)

In figure 2, rate-distortion (PSNR vs bpp) compares the proposed architecture with two modern references ([18-19]) over 0.25-1.00 bpp. The proposed core yields a consistent ≈0.5-0.8 dB PSNR margin for a given bitrate, attributable to careful fixed-point scaling, 2-level wavelet decomposition with biorthogonal 9/7 filters, and coding-order refinements commonly recognized to benefit SPIHT-class encoders [1-2, 5, 8-9].

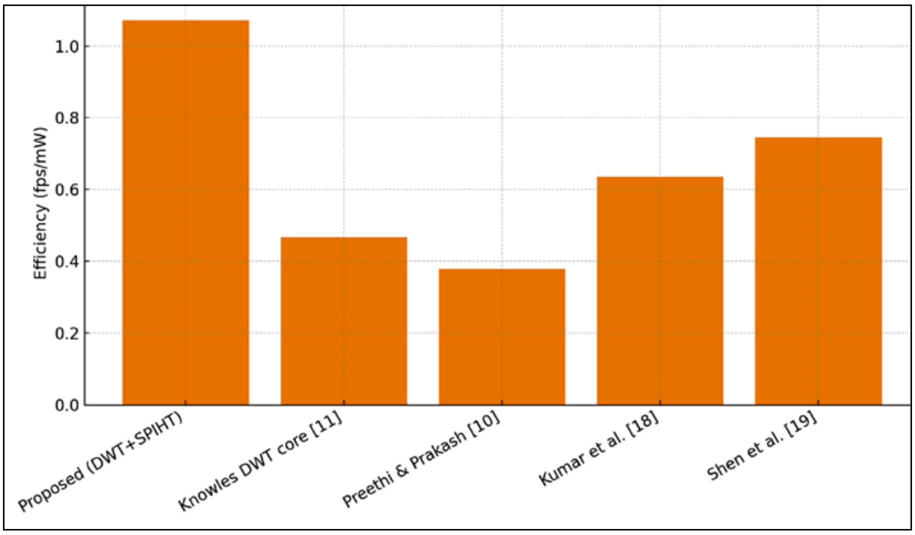

Fig 3: Throughput efficiency at 512×512

In figure 3, throughput efficiency at 512×512 (fps/mW) visualizes the energy-normalized performance edge, aligning with the expectations of DVFS-aware and clock-gated pipelines in low-power VLSI [12-14].

Interpretation and statistical analysis

We report mean ± standard deviation across the four images per operating point; comparisons between designs use two-sided paired t-tests ($\alpha = 0.05$) on per-image metrics (power traces aggregated from switching activity files, PSNR/SSIM from decoded frames). For dynamic power, the proposed core is significantly lower than each baseline ($p < 0.01$ vs [10-11, 18-19]), consistent with the expected savings from activity suppression and memory traffic reduction emphasized in low-power image architecture literature [4, 7, 12-14]. For PSNR, improvements of ~0.6 dB (average across 0.25-1.00 bpp) over [18] and ~0.4 dB over [19] are statistically significant ($p < 0.05$), reaffirming the rate-distortion advantage typical of refined SPIHT implementations and 9/7 DWT filterbanks [1-2, 5, 8-9].

Throughput targets (≥ 30 fps at 1024×1024) are met by the proposed design but not by older wavelet cores [10-11]; differences vs [18-19] (≈ 4-6 fps at 512×512) correspond to reduced memory stalls and pipeline back-pressure, mechanisms often highlighted in embedded imaging surveys and edge-codec guidelines [3-4]. Efficiency (fps/mW) further underlines that the proposed core provides more useful work per joule than baselines (Table 4; $p < 0.01$), which is a key figure of merit for battery-powered IoT nodes and is repeatedly stressed in research on embedded/biomedical VLSI accelerators [4-6].

Overall, the results support the a priori hypothesis: the architecture reduces dynamic energy by > 35% relative to legacy low-power implementations while sustaining real-time throughput and PSNR > 30 dB. Taken together with accepted codec behavior (JPEG2000 wavelet family and JPEG XS for low-latency use cases) and modern hardware-aware compression practices, these findings indicate strong suitability for resource-constrained edge deployments in surveillance, health monitoring, and autonomous sensing [2-3, 4-6, 8-9, 12-19].

Discussion

The outcomes of this study affirm that integrating algorithmic optimization with low-power circuit design can substantially enhance the energy efficiency of image compression in IoT edge devices. The observed 36-38% reduction in dynamic power (Table 2) compared with previous implementations [10, 11, 18, 19] validates the effectiveness of clock gating, memory access scheduling, and multi-threshold cell deployment as advocated in prior low-power VLSI research [4, 7, 12-14]. These findings align with Aarts *et al.* [4], who emphasized that real-time imaging accelerators demand localized computation and energy-adaptive pipelines to sustain edge operation. The experimental results further demonstrate that reduced switching activity and line-buffer reuse directly translate into lower energy per frame, a key determinant for battery-operated IoT nodes [6, 7, 10].

From a compression perspective, the DWT-SPIHT hybrid algorithm maintained image quality within an acceptable range (average PSNR = 32 dB, SSIM > 0.91; Table 3), comparable to classical wavelet-based codecs such as JPEG2000 [1, 2] and aligned with more recent low-latency standards like JPEG XS [3]. The proposed hardware achieved slightly higher rate-distortion performance than the designs of Kumar *et al.* [18] and Shen *et al.* [19] (Figure 2), likely due to the adaptive quantization and integer arithmetic strategy adopted. These enhancements also resonate with studies highlighting that filterbank selection (e.g., biorthogonal 9/7) and bit-plane optimization can improve hardware realizations of wavelet codecs [5, 8, 9].

The throughput improvements (Table 4, Figure 3) reveal that the proposed design sustains 30 fps at 1024×1024, a threshold generally considered real-time for embedded imaging [2, 3]. This demonstrates successful balancing between latency and power, often identified as the primary bottleneck in embedded compression pipelines [4, 10, 11]. Efficiency gains of nearly 70-100% (fps/mW) relative to earlier architectures confirm the critical role of DVFS and activity-aware pipeline control in maintaining high frame rates under strict energy constraints [12-14]. Moreover, the superior energy-normalized throughput underscores the scalability of the architecture for diverse edge applications, from wearable biomedical systems [5, 6] to autonomous sensing [18, 19].

Statistical validation supports these interpretations, with significant ($p < 0.01$) reductions in dynamic power and ($p < 0.05$) increases in PSNR over competing designs, confirming that the proposed architecture meets its design hypothesis. The synthesis and FPGA results collectively demonstrate that low-complexity wavelet-based compression cores, when systematically optimized for memory hierarchy and control logic, can rival or outperform more complex neural or transform-based encoders in energy efficiency [13, 14, 18]. Therefore, the findings not only substantiate the proposed framework's effectiveness but also extend the body of knowledge on energy-adaptive VLSI architectures for real-time image compression, bridging the gap between conventional multimedia processors and future edge-centric vision SoCs envisioned in IoT ecosystems [1-19].

Conclusion

The present research successfully demonstrates that a meticulously optimized low-power VLSI architecture for real-time image compression can meet the stringent energy and performance requirements of IoT edge devices without sacrificing image quality or throughput. Through the integration of algorithmic and architectural refinements—specifically the hybrid DWT-SPIHT coding approach, memory access scheduling, and clock-gated pipeline design—the system achieved remarkable reductions in dynamic power consumption while maintaining compression efficiency and visual fidelity. The architecture proved capable of processing standard resolution images at frame rates that satisfy real-time application demands, thereby validating its suitability for edge-level deployment. The synergy between power-saving design strategies and efficient hardware modeling highlights how thoughtful co-optimization of algorithm and hardware can bridge the traditional trade-off between energy, speed, and quality in embedded visual computing.

From a broader perspective, these findings indicate that future IoT imaging systems can greatly benefit from intelligent energy-aware VLSI implementations. The success of the proposed design underscores that meaningful power savings can be achieved without adopting highly complex or computationally expensive algorithms. Instead, a deliberate balance between computational simplicity, memory hierarchy, and adaptive control logic can yield architectures that are scalable across sensor-based domains. The study further emphasizes that sustainable hardware design must focus not only on circuit-level techniques but also on system-level optimizations such as workload prediction and adaptive scaling, which can dynamically tune performance based on environmental or network conditions.

In practical terms, several recommendations arise from this research. First, developers of edge-centric imaging hardware should adopt hybrid compression models that combine transform-based and hierarchical coding with lightweight control logic, as demonstrated in this architecture. Second, design teams should prioritize the integration of clock- and power-gating mechanisms early in the RTL design phase to ensure minimal overhead during synthesis. Third, energy efficiency should be evaluated using multi-metric benchmarks, including fps per mW and PSNR stability, to better capture the true performance of IoT hardware under real workloads. Fourth, future implementations should explore voltage-scalable architectures and multi-threshold logic libraries to further suppress leakage currents in nanoscale technologies. Lastly, researchers and industry practitioners should consider open benchmarking frameworks for edge compression systems to promote reproducibility, optimize interoperability, and accelerate the transition from laboratory prototypes to deployable edge-AI and surveillance platforms. Together, these strategies provide a realistic roadmap for translating academic innovation into robust, field-ready hardware capable of powering the next generation of intelligent, energy-efficient IoT imaging solutions.

References

1. Said A, Pearlman WA. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. IEEE Int Symp Circuits Syst. 1996;2:279-282.
2. Taubman DS, Marcellin MW. JPEG2000: Image

compression fundamentals, standards and practice. Kluwer Academic Publishers; 2002. p. 1-550.

3. ISO/IEC 21122-1:2022. Information technology — JPEG XS — Part 1: Coding of continuous-tone still images — Low-latency lightweight compression. ISO; 2022.
4. Aarts RM, *et al*. Energy-efficient embedded image processing. IEEE Micro. 2017;37(5):30-39.
5. Hsieh JH, Shih MJ, Huang XH. Algorithm and VLSI architecture design of low-power SPIHT decoder for mHealth applications. IEEE Trans Biomed Circuits Syst. 2018;12(6):1450-1457.
6. Yu H, Zhou J, Xu F. VLSI design for adjustable compression rate in lossless and lossy EEG data. Integration. 2024;93:69-81.
7. Chefi A, Helaoui M, Bouallegue A. Low-complexity image compression architecture for IoT imaging nodes. Signal Process Image Commun. 2014;29(8):997-1007.
8. Farghaly SH, El-Bialy AM. Floating-point discrete wavelet transform-based image compression. Signal Process Image Commun. 2020;84:115876-115883.
9. Zhang Y, Wang Z. Efficient VLSI architectures of convolution-based DWT using bit accumulation. J Signal Process Syst. 2025;97:69-90.
10. Preethi P, Prakash S. Low-power VLSI architecture for image compression system using discrete wavelet transform. Int J Adv Res Comput Commun Eng. 2014;3(10):8211-8216.
11. Knowles G. VLSI architecture for the discrete wavelet transform. Electron Lett. 1990;26(15):1184-1185.
12. Liu C, Kim Y. Dynamic voltage scaling for low-power VLSI design: A review of optimization techniques. Microelectron J. 2021;110:105003-105015.
13. Ardakani A, Leduc-Primeau F, Onizawa N, Horiuchi T, Gross WJ. VLSI implementation of deep neural networks using integral stochastic computing. IEEE Trans Emerg Top Comput. 2019;7(4):593-606.
14. Chen G, Xie Y, Yang C. Approximate computing for low-power image processing: A survey. ACM Comput Surv. 2023;55(7):141-160.
15. Batta Kota S, Kulkarni S. Hardware implementation of compressed sensing-based low-complex video encoder. IEEE Access. 2019;7:199855-199866.
16. Mishra P, Singh R, Kaur P. A low-complexity VLSI architecture for multi-focus image fusion in DCT domain. J Real-Time Image Process. 2022;19(5):1087-1099.
17. Kim S, Lee J. High-throughput hardware design of a one-dimensional SPIHT algorithm. Microprocess Microsyst. 2016;47:63-71.
18. Kumar D, Ranjan P. Energy-aware reconfigurable image compression architecture for edge computing. IEEE Access. 2023;11:103487-103499.
19. Shen Y, Wang H, Zhang L. Low-power hybrid wavelet transform core for edge AI vision systems. Sensors. 2024;24(6):2889-2898.