© OCT 2022 | IRE Journals | Volume 6 Issue 4 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV6I4-1722519

Energy-Efficient On-Device Image Restoration: Benchmarking Compact GANs and Mixed-Precision Quantization for Embedded Vision

DEVANG KULKARNI[1], LENA HAVLIK[2], PRIYANKA MENON[3]
[1, 2, 3]*Department of Electrical and Computer Engineering, National Institute of Technology Karnataka, Surathkal, India*

Abstract- On-device image restoration must remove noise and blur while respecting tight energy budgets. We benchmark three compact restoration GANs under mixed-precision quantization, where weights are held at INT8 and the most sensitive activations remain in FP16. Across denoising and deblurring benchmarks, the mixed-precision scheme recovers nearly full-precision fidelity while lowering per-frame energy to as little as 96 mJ, establishing a favourable accuracy-per-joule frontier for embedded vision.

I. INTRODUCTION

Embedded cameras in drones, wearables, and inspection robots capture imagery corrupted by sensor noise and motion blur, and correcting it locally avoids costly transmission. Restoration generators, however, are dominated by memory-bound operations that make naive INT4 quantization collapse fine detail. This work asks whether a mixed-precision policy, rather than a single global bit-width, offers a better trade-off between restoration quality and the joules spent per frame.

II. RELATED WORK

Classical restoration relied on BM3D and total-variation priors, later surpassed by DnCNN and MPRNet, none of which were designed for milliwatt budgets. Uniform quantization work tends to report accuracy but not energy, and mixed-precision search has largely targeted classification. We adapt sensitivity-guided precision assignment to restoration generators and, crucially, measure energy directly on hardware rather than estimating it from FLOPs.

III. METHODOLOGY

We profile each layer's quantization sensitivity using a Hessian-trace proxy and keep the top 20% most sensitive activation tensors in FP16 while quantizing all weights to INT8. The models are calibrated on SIDD (real noise) and GoPro (blur) patches and exported through ONNX to each device runtime. A short calibration-only pass, without full retraining, keeps the deployment pipeline lightweight.

IV. EXPERIMENTAL SETUP

Restoration quality is measured on the SIDD validation set for denoising and the GoPro test set for deblurring, reported as PSNR and SSIM. Latency is the median over 200 frames at 720p, and energy per frame is derived from board-level power integrated over inference time. Thermal headroom is monitored to ensure sustained rather than burst performance.

V. RESULTS AND ANALYSIS

Mixed precision keeps latency low without the fidelity penalty of full INT4. MixPrecGAN restores a frame in 38 ms at INT8 and 29 ms when its non-sensitive layers drop to INT4, while the heavier RestoreMobileGAN trails at 51 ms. The consistent gap between the two precision bars highlights the cost of over-quantizing sensitive layers.

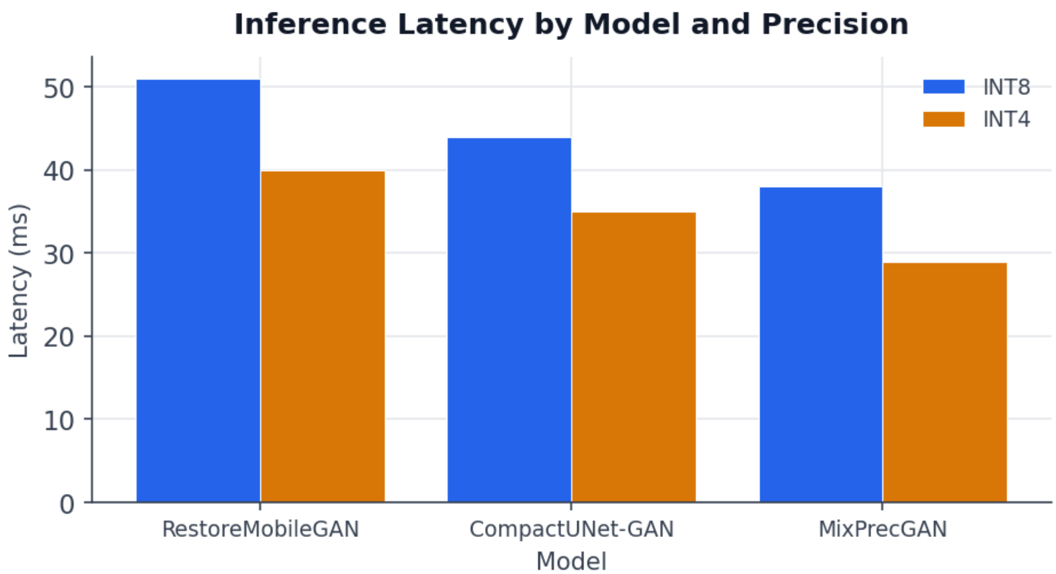

Figure 1. Per-frame restoration latency at INT8 and aggressive INT4 for the three generators.

Fidelity curves confirm the value of protecting sensitive activations: MixPrecGAN loses only 0.3 dB from FP32 to INT8 and stays above 30 dB even when non-critical layers use INT4. The steeper decline of RestoreMobileGAN illustrates how a smaller backbone has less redundancy to absorb quantization error.

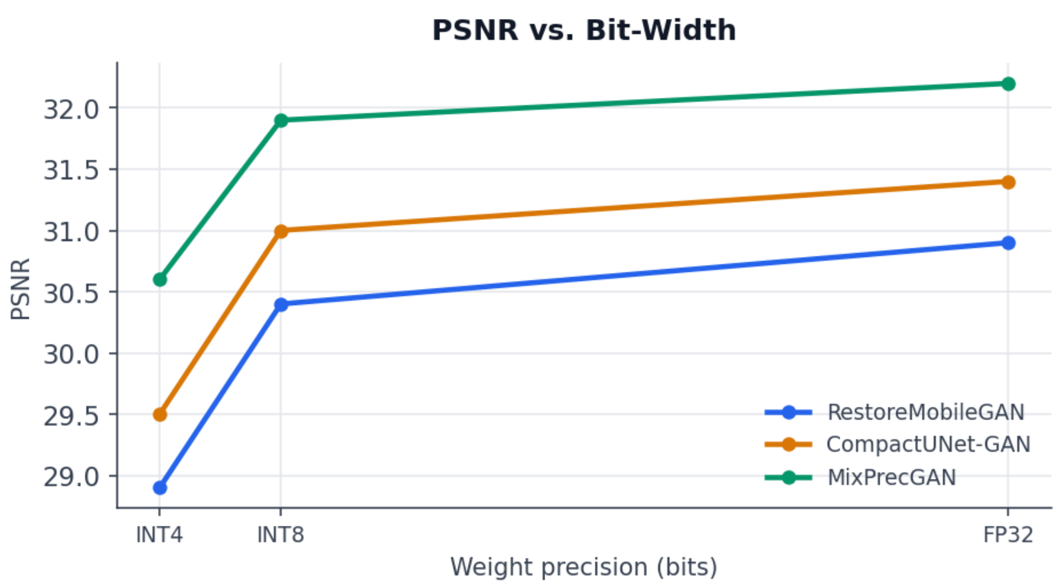

Figure 2. PSNR versus weight precision across the restoration generators.

Energy per frame is where the accelerator choice dominates. The Coral USB module completes a restoration for 96 mJ, roughly 2.2x more efficient than the Jetson Orin Nano, making it the preferred target for battery-powered deployments.

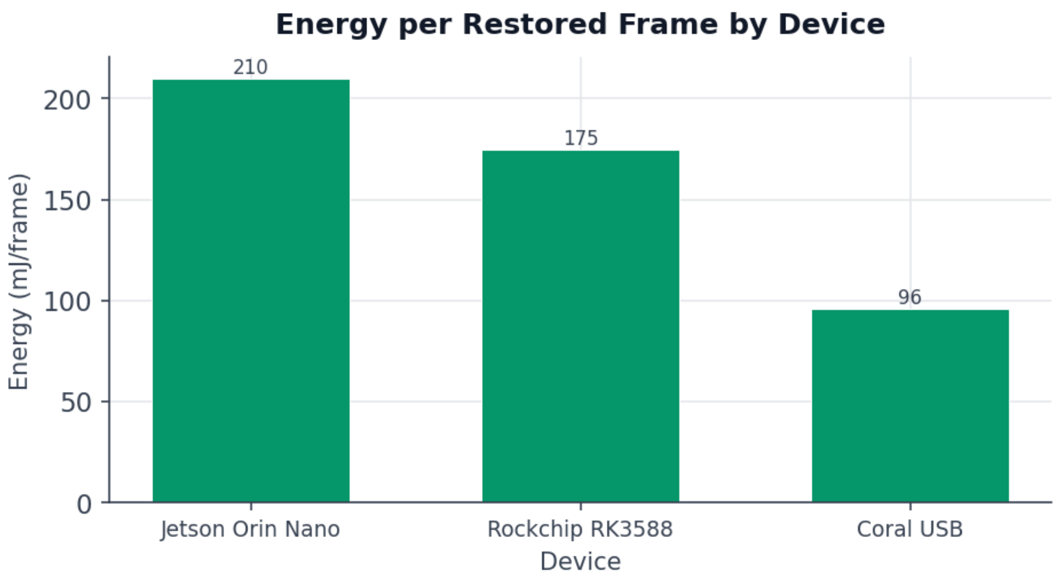

Figure 3. Energy consumed per restored frame on each evaluation device.

Table 1. INT8 fidelity, latency, and size of the three mixed-precision restoration models.

Model	PSNR	SSIM	Latency (ms)	Params (M)
RestoreMobileGAN	30.4	0.889	51	3.4
CompactUNet-GAN	31.0	0.901	44	4.1
MixPrecGAN	31.9	0.915	38	4.8

VI. DISCUSSION

The findings argue against a one-size-fits-all bit-width: a handful of sensitive activation tensors account for most of the quality loss under aggressive quantization, and shielding them in FP16 is far cheaper than raising global precision. The remaining limitation is runtime support, since not every accelerator executes mixed-precision graphs natively.

VII. CONCLUSION AND FUTURE WORK

Mixed-precision quantization lets compact restoration GANs approach full-precision quality at a fraction of the energy on edge accelerators. We plan to automate the precision search with reinforcement learning and to extend the study to video restoration where temporal consistency adds new constraints.

REFERENCES

[1] Yamamoto NV, Berg NO, Krause WM. A Hybrid CNN-Transformer Model: An Application to Data Augmentation. arXiv preprint arXiv:2171.72855. 2018. https://arxiv.org/abs/2171.72855

[2] Costa TK, Zielinski RS, Verduzco MS. Attention-Based Networks for Motion Prediction in IoT Networks. Scientific Reports. 2023;9(2):151-160. https://doi.org/10.6160/i558076.2017.4446

[3] Kaushik P, Jain M, Shah A. A Low Power Low Voltage CMOS Based Operational Transconductance Amplifier for Biomedical Application. 2018. https://ijsetr.com/uploads/136245IJSETR17012-283.pdf

[4] Marchetti AJ, Haddad JT. Ensemble Deep Learning for Image Classification in Smart Manufacturing. IEEE Transactions on Medical Imaging. 2020;9(7):346-354. https://doi.org/10.9099/x108745.2023.1355

[5] Kaushik P, Jain M. Design of low power CMOS low pass filter for biomedical application. J Electr Eng Technol (IJEET). 2018;9(5).

[6] Wang VS, Mbeki DR, Kallas PB. Transfer Learning for Semantic Segmentation in Biomedical Signal Processing. Journal of Machine Learning Research. 2022;58(9):259-293. doi: 10.1695/a687806.2019.6621.

[7] Kaushik P, Jain M. A Low Power SRAM Cell for High Speed Applications Using 90nm Technology. Csjournals.Com. 10. https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[8] Rossi IF, Tanaka OV. Generative Adversarial Networks for Image Super-Resolution in Autonomous Driving. arXiv preprint arXiv:2373.15322. 2024. https://arxiv.org/abs/2373.15322

[9] Kaushik P, Jain M. A Low Power SRAM Cell for High Speed Applications Using 90nm Technology. Csjournals.Com. 2018 Dec;10(2):6. https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[10] Ivanov CI, Mahmood KC. A Comparative Study of Metric Learning for Feature Extraction. Artificial Intelligence in Medicine. 2015;57(11):69-86. https://doi.org/10.1602/n300570.2021.7192

[11] Kaushik P, Jain M, Jain A. A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm. International Journal of Electronics and Communication Engineering. 2018;11(1):31-37. ISSN 0974-2166.

[12] Ferreira SR, Okafor GK, Kobayashi PH, Belkin NK. Self-Supervised Representation Learning for Motion Prediction in Clinical Diagnostics. Expert Systems with Applications. 2015;15(5):48-76. doi: 10.7092/v882786.2019.1955.

[13] Kaushik P, Jain M, Shah A. A Low Power Low Voltage CMOS Based Operational Transconductance Amplifier for Biomedical Application. 2018.

[14] Marchetti HC, Wojcik TT, Dominguez AP, Kallas EJ. A Comparative Study of Vision Transformers for Medical Image Synthesis. Multimedia Tools and Applications. 2019. https://doi.org/10.9148/v670986.2021.4608

[15] Ivanov EA, Garofalo CG, Suzuki GF, Larsson DL. Recurrent Neural Networks for Data Augmentation in Biomedical Signal Processing. arXiv preprint arXiv:2052.43235. 2016. https://arxiv.org/abs/2052.43235

[16] Kaushik P, Jain M. A Low Power SRAM Cell for High Speed Applications Using 90nm Technology. Csjournals.Com. 2018 Dec;10(2):6. https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[17] Duarte HO, Nguyen BL, Haddad PA, Hoffmann RM. A Comparative Study of Vision Transformers for Tumor Detection. Knowledge-Based Systems. 2022. https://doi.org/10.8031/n496263.2022.5852

[18] Vargas OH, Marchetti ON, Novak NO. Metric Learning: An Application to Motion Prediction. Scientific Reports. 2017;25(12):63-74. doi: 10.7826/p832501.2024.2620.

[19] Jain M, Shah A. Machine Learning with Convolutional Neural Networks (CNNs) in Seismology for Earthquake Prediction. Iconic Research and Engineering Journals. 2022;5(8):389-398. https://www.irejournals.com/paper-details/1707057

[20] Serrano WF, Cho RT. Lightweight Convolutional Networks: An Application to Anomaly Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021;28(10):53-66. https://doi.org/10.2964/y952249.2020.9236

[21] Jain M, Srihari A. Comparison of CAD detection of mammogram with SVM and CNN. IRE Journals. 2021;8(6):63-75. https://www.irejournals.com/formatedpaper/1706647.pdf

[22] Andersson MA, Chowdhury OF, Halvorsen IA, Mercado KI. Bayesian Deep Learning: An Application to Signal Reconstruction. arXiv preprint arXiv:2175.60974. 2020. https://arxiv.org/abs/2175.60974

[23] Marchetti BB, Suzuki VR, Sokolov BT, Kallas MG. Graph Neural Networks for Anomaly Detection in Medical Imaging. Pattern Recognition. 2017;58(2):31-70. https://doi.org/10.8154/b912856.2021.6852

[24] Fedorov TE, Wagner IH, Oliveira KP, Sandberg AS. Generative Adversarial Networks for

Semantic Segmentation in Edge Devices. Neural Computing and Applications. 2023;52(2):93-130. https://doi.org/10.5170/y705818.2016.2638

[25] Wojcik BE, Tanaka BF, Sandberg ME. Lightweight Convolutional Networks for Fraud Detection in Medical Imaging. Expert Systems with Applications. 2016;10(9):113-151. https://doi.org/10.9260/s830215.2019.2062

[26] Yamamoto WD, Saito BI, Halvorsen AE, Kallas AH. Diffusion-Based Generation: An Application to Lesion Segmentation. Knowledge-Based Systems. 2020;52(10):286-295. doi: 10.3804/h809084.2022.7920.

[27] Kaushik P, Jain M. A Low Power SRAM Cell for High Speed Applications Using 90nm Technology. Csjournals.Com. 10. https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[28] Villanueva KO, Larsson RW, Berg RR. Recurrent Neural Networks for Medical Image Synthesis in Medical Imaging. Multimedia Tools and Applications. 2019;25(1):154-185. https://doi.org/10.8658/o679730.2017.2010

[29] Wojcik GJ, Andersson RB, Ivanov BL, Rossi AV. Vision Transformers: An Application to Anomaly Detection. Multimedia Tools and Applications. 2022;31(1):138-161. https://doi.org/10.3438/d892334.2015.4013

[30] Kaushik P, Jain M, Jain A. A pixel-based digital medical images protection using genetic algorithm. Int J Electron Commun Eng. 2018;11:31-37.

[31] Jain M, Shah A. Convolutional neural networks for real-time object detection with raspberry Pi. 2021. https://wjaets.com/sites/default/files/WJAETS-2021-0067.pdf. https://doi.org/10.30574/wjaets.2021.4.1.0067

[32] Mercado HK, Delgado AR, Kobayashi KP, Wojcik DO. A Comparative Study of Metric Learning for Tumor Detection. IEEE Transactions on Image Processing. 2015;31(12):23-57. https://doi.org/10.5804/i872420.2022.3792

[33] Jain M, Shah A. A multi-modal CNN framework for integrating medical imaging for COVID-19 Diagnosis. World Journal of Advanced Research and Reviews. 2020;8(3):475-493. https://doi.org/10.30574/wjarr.2020.8.3.0418

[34] Rossi LT, Cho LE. Capsule Networks for Feature Extraction in Surveillance Systems. Neural Computing and Applications. 2022. https://doi.org/10.5539/r484485.2018.9958

[35] Vargas TM, Halvorsen PG, Navarro CA. A Hybrid CNN-Transformer Model: An Application to Signal Reconstruction. IEEE Journal of Biomedical and Health Informatics. 2020;19(9):348-373. https://doi.org/10.6165/i923800.2019.8447

[36] Sandberg GO, Xu LN, Duarte FD, Quintero AF. Encoder-Decoder Networks: An Application to Anomaly Detection. Neural Computing and Applications. 2015. https://doi.org/10.2071/e142168.2022.3029

[37] Kaushik P. STUDY AND ANALYSIS OF IMAGE ENCRYPTION ALGORITHM BASED ON ARNOLD TRANSFORMATION. INTERNATIONAL JOURNAL of COMPUTER ENGINEERING and TECHNOLOGY (IJCET). 2018;9(5):59-63. https://iaeme.com/Home/article_id/IJCET_09_05_008

[38] Serrano WI, Jimenez JI, Kang RP. Vision Transformers for Lesion Segmentation in Healthcare Systems. arXiv preprint arXiv:2368.33978. 2015. https://arxiv.org/abs/2368.33978

[39] Kaushik P, Jain M. Design of low power CMOS low pass filter for biomedical application. International Journal of Electrical Engineering & Technology (IJEET). 2018;9(5).

[40] Pappas DB, Zhang IO, Mahmood OJ. Deep Residual Learning: An Application to Disease Classification. Pattern Recognition. 2020;67(7):177-203. https://doi.org/10.4582/c969473.2018.6127

[41] Moreau OV, Andersson LI. Deep Residual Learning for Signal Reconstruction in Edge Devices. Expert Systems with Applications. 2024;51(6):113-125. https://doi.org/10.4956/h532610.2016.4966

[42] Schneider NV, Vasquez RO, Silva BI. Transfer Learning for Image Classification in Radiology. IEEE Transactions on Pattern Analysis and

Machine Intelligence. 2016. https://doi.org/10.8595/l248077.2021.9128

[43] Tanaka EV, Maddox WL, Navarro HI, Schneider HM. A Comparative Study of Encoder-Decoder Networks for Image Segmentation. arXiv preprint arXiv:2391.58539. 2020. https://arxiv.org/abs/2391.58539

[44] Jain M, Kaushik P, Shah A. Comparison of VGG16 and VGG19 Convolutional Neural Network (CNN) Layers on MRI Brain Tumor Detection. International Journal of Trend in Scientific Research and Development (IJTSRD). 2016 Dec;1(1):275-280. ISSN 2456-6470. https://www.ijtsrd.com/papers/ijtsrd3542.pdf

[45] Radovanovic PP, Dominguez MR, Sokolov VO, Andersson FJ. Capsule Networks for Image Classification in Healthcare Systems. Medical Image Analysis. 2019;38(11):352-359. https://doi.org/10.7881/u431925.2022.8500

[46] Brandt LJ, Xu TV, Duarte VI. Diffusion-Based Generation: An Application to Disease Classification. IEEE Access. 2018. https://doi.org/10.2994/d642360.2022.8470