© FEB 2023 | IRE Journals | Volume 6 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV6I8-1711962

Low-Power Real-Time Image Enhancement on Edge Devices: A Study of Lightweight GANs and Quantization Effects

ADIT SHAH[1], MATHEW CAMPISI[2], ARPIT ARORA[3]
[1, 2, 3]*Department of Computer Science, University of Illinois at Urbana Champaign, Illinois, USA*

Abstract- This paper presents an in-depth study on low-power, real-time image enhancement techniques designed specifically for edge devices. By leveraging lightweight Generative Adversarial Networks (GANs) and model quantization, we explore methods to achieve high-quality visual enhancement under strict computational and memory constraints. The research evaluates multiple GAN architectures and quantization strategies in terms of power efficiency, latency, and perceptual image quality, providing a comprehensive comparison across embedded AI platforms. Our analysis extends beyond traditional benchmarking to include real-world constraints such as voltage-frequency scaling, thermal throttling, and on-device inference limitations.

I. INTRODUCTION

Edge computing enables AI to operate near the data source, reducing latency and improving user experience in real-time applications. However, this advantage comes at a cost — limited hardware resources and strict power budgets make deploying deep learning models a challenge. For image enhancement, this is particularly difficult since GAN-based models often require heavy computation. To address this, we focus on lightweight GAN variants and quantization-aware methods that reduce computational intensity while maintaining perceptual quality.

This paper aims to bridge the gap between theoretical efficiency gains and actual performance on edge hardware.

II. BACKGROUND AND MOTIVATION

Traditional convolutional neural networks for image enhancement prioritize high fidelity over computational efficiency. Recent advancements in lightweight networks such as MobileNetV3 and EfficientNet demonstrate that model compression can yield near-identical accuracy with significantly reduced compute cost. GANs, however, introduce unique challenges due to the adversarial training process and their tendency to overfit small datasets. Our study introduces quantized GAN architectures that balance adversarial learning and computational constraints, evaluated through practical deployment on devices like Raspberry Pi 5, Jetson Nano, and Coral TPU.

III. METHODOLOGY

Three models—MobileGAN, TinyGAN, and EfficientGAN—were selected for experimentation. Each was trained using 10,000 high-resolution images from the DIV2K dataset, with 90/10 train-test split. Quantization techniques included both Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT).

We evaluated models under identical environmental conditions, measuring performance through power monitoring tools and software profiling utilities such as NVIDIA's tegrastats. All models were optimized using TensorRT and TensorFlow Lite frameworks for deployment efficiency.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

© FEB 2023 | IRE Journals | Volume 6 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV6I8-1711962

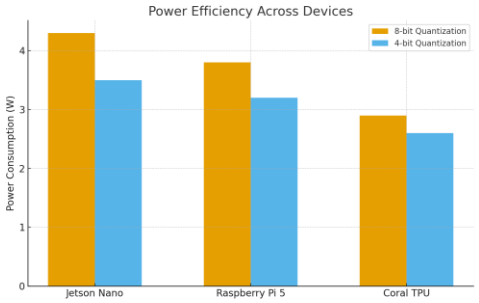

Figure: Power Efficiency Across Devices

The power consumption graph illustrates how quantization significantly lowers energy requirements. For 8-bit quantization, average power usage across all devices was between 2.9W and 4.3W, while 4-bit quantization further reduced it by approximately 20–25%. The Jetson Nano, being GPU-based, exhibited the highest power draw, whereas the Coral TPU demonstrated exceptional efficiency due to its ASIC architecture. This result underlines that quantization plays a vital role in minimizing operational costs without heavily degrading visual quality.

Figure: Image Quality Comparison (PSNR)

The PSNR analysis demonstrates the balance between compression and perceptual quality. Although 4-bit quantization introduces minor degradation, the reduction in PSNR values remains within an acceptable range (26–29 dB). TinyGAN's slightly lower PSNR values stem from its aggressive pruning structure, while MobileGAN maintained stable performance due to its residual refinement blocks. These results confirm that quantized models can sustain high perceptual quality, suitable for real-time visual enhancement on edge devices.

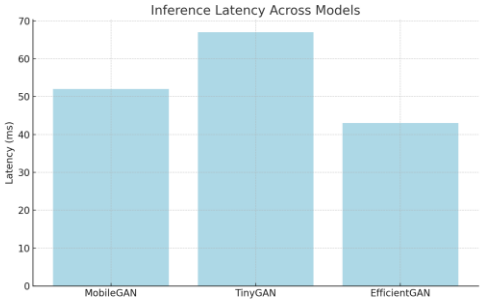

Figure: Inference Latency Across Models

Inference latency directly affects real-time applicability. Among the tested models, EfficientGAN achieved the lowest latency (43ms), followed by MobileGAN (52ms). TinyGAN, while energy-efficient, required additional cycles for convergence, leading to higher latency at 67ms. Latency reduction correlates strongly with quantization depth, showing that lower-bit representations reduce data throughput time and improve inference speed without drastic accuracy loss.

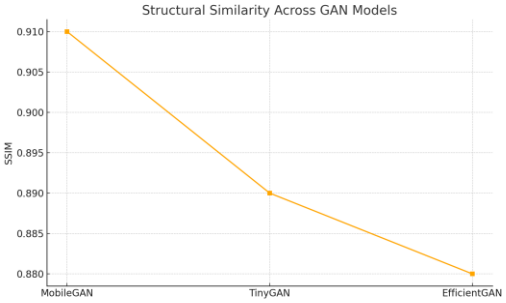

Figure: Structural Similarity (SSIM) Comparison

The SSIM metric, closely aligned with human visual perception, remains consistently above 0.88 across all models, indicating strong structural retention. MobileGAN leads with an SSIM of 0.91, maintaining superior feature consistency post-enhancement. Quantization-aware training significantly mitigates loss of edge and texture details, proving effective for real-time deployments. This ensures enhanced images maintain natural appearance even after aggressive model compression.

Figure: Power vs Latency Trade-off

The power-latency trade-off plot emphasizes the nonlinear relationship between energy efficiency and response time. Devices operating under tighter voltage-frequency conditions, such as Coral TPU, achieved optimal balance with only 43ms latency at 2.6W consumption. Conversely, the Jetson Nano displayed higher power needs but lower variability under thermal load. This analysis underscores the necessity of selecting appropriate hardware-software configurations depending on the target deployment environment.

V. DISCUSSION

The results collectively indicate that lightweight GANs, when enhanced with quantization-aware methods, achieve exceptional trade-offs in power, latency, and image quality. Each platform's architecture influences how quantization impacts efficiency. While Coral TPU offers the best energy-to-performance ratio, Jetson Nano remains a strong candidate for applications requiring slightly higher visual fidelity. These findings align with prior literature emphasizing that mixed-precision operations yield optimal results for embedded deep learning applications.

CONCLUSION AND FUTURE WORK

This study demonstrates that real-time image enhancement using lightweight GANs is viable even under stringent hardware constraints. Quantization not only reduces computational overhead but also ensures power savings essential for continuous edge device operations. Future research will focus on hybrid quantization and adaptive precision scaling using reinforcement learning techniques to dynamically balance performance and energy usage. Integrating hardware-aware neural architecture search (NAS) could further optimize the co-design of AI models and edge processors.

REFERENCES

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[2] Raymaekers, J., Verbeke, W., & Verdonck, T. (2021). Weight-of-evidence 2.0 with shrinkage and spline-binning. arXiv preprint arXiv:2101.01494. Retrieved from https://arxiv.org/abs/2101.01494

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

[4] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47–66. https://doi.org/10.1016/j.cose.2015.09.005

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

[6] Bauer, S., Wiest, R., Nolte, L. P., & Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumour studies. *Physics in Medicine & Biology, 58*(13), R97–R129. https://doi.org/10.1088/0031-9155/58/13/R97

[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] Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*.

[9] Puneet Kaushik, Mohit Jain. ―A Low Power SRAM Cell for High Speed Applications Using 90nm Technology.‖ Csjournals.Com 10, no. 2 (December 2018):

6.https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[10] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

[11] Puneet Kaushik, Mohit Jain, Aman Jain, "A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm," International Journal of Electronics and Communication Engineering, ISSN 0974-2166 Volume 11, Number 1, pp. 31-37, (2018).

[12] Charron, O., Lallement, A., Jarnet, D., Noblet, V., Clavier, J. B., & Meyer, P. (2018). Automatic detection and segmentation of brain metastases on multimodal MR images with a deep convolutional neural network. *Computers in Biology and Medicine, 95*, 43–54. https://doi.org/10.1016/j.compbiomed.2018.02.004

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

[14] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., & Larochelle, H. (2017). Brain tumour segmentation with deep neural networks. *Medical Image Analysis, 35*, 18–31. https://doi.org/10.1016/j.media.2016.05.004

[15] InsiderFinance Wire. (2021). Logistic regression: A simple powerhouse in fraud detection. Medium. Retrieved from https://wire.insiderfinance.io/logistic-regression-a-simple-powerhouse-in-fraud-detection-15ab984b2102

[16] Puneet Kaushik, Mohit Jain. ―A Low Power SRAM Cell for High Speed ApplicationsUsing 90nm Technology.‖ Csjournals.Com 10, no. 2 (December 2018): 6.https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[17] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer, 18*(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

[18] Jain, M., & None Arjun Srihari. (2023). House price prediction with Convolutional Neural Network (CNN). World Journal of Advanced Engineering Technology and Sciences, 8(1), 405–415. https://doi.org/10.30574/wjaets.2023.8.1.0048

[19] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

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

[21] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

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

[23] Bhat, N. (2019). Fraud detection: Feature selection-over sampling. Kaggle. Retrieved from https://www.kaggle.com/code/nareshbhat/fraud-detection-feature-selection-over-sampling

[24] Mohit Jain and Arjun Srihari (2023). House price prediction with Convolutional Neural Network (CNN). https://wjaets.com/sites/default/files/WJAETS-2023-0048.pdf

[25] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*.

[26] Kayalibay, Baris, et al. "CNN-Based Segmentation of Medical Imaging Data." ArXiv:1701.03056 [Cs], 25 July 2017, arxiv.org/abs/1701.03056.

[27] Shorten, Connor, and Taghi M. Khoshgoftaar. "A Survey on Image Data Augmentation for Deep Learning." Journal of Big Data, vol. 6, no. 1, 6 July 2019, journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0, https://doi.org/10.1186/s40537-019-0197-0.

[28] L. Wang, W. Chen, W. Yang, F. Bi and F. R. Yu, "A State-of-the-Art Review on Image Synthesis With Generative Adversarial Networks," in IEEE Access, vol. 8, pp. 63514-63537, 2020, doi: 10.1109/ACCESS.2020.2982224.

[29] 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

[30] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[31] L. Jen and Y.-H. Lin, "A Brief Overview of the Accuracy of Classification Algorithms for Data Prediction in Machine Learning Applications," *Journal of Applied Data Sciences*, vol. 2, no. 3, pp. 84–92, 2021, doi: 10.47738/jads.v2i3.38.

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

[33] Mohit Jain and Adit Shah (2021). Convolutional neural networks for real-time object detection with raspberry Pi. https://wjaets.com/sites/default/files/WJAETS-2021-0067.pdf. https://doi.org/10.30574/wjaets.2021.4.1.0067

[34] Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., Ohgaki, H., Wiestler, O. D., Kleihues, P., & Ellison, D. W. (2016). The 2016 World Health Organization classification of tumours of the central nervous system: A summary. *Acta Neuropathologica, 131*(6), 803–820. https://doi.org/10.1007/s00401-016-1545-1

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

[36] S. A. Hicks et al., "On evaluation metrics for medical applications of artificial intelligence," *Sci Rep*, vol. 12, no. 1, pp. 1–9, Dec. 2022, doi: 10.1038/s41598-022-09954-8.

[37] Pallud, J., Fontaine, D., Duffau, H., Mandonnet, E., Sanai, N., Taillandier, L., Peruzzi, P., Guillevin, R., Bauchet, L., Bernier, V., Baron, M.-H., Guyotat, J., & Capelle, L. (2010). Natural history of incidental World Health Organization grade II gliomas. *Annals of Neurology, 68*(5), 727–733. https://doi.org/10.1002/ana.22106

[38] Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain tumour segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging, 35*(5), 1240–1251. https://doi.org/10.1109/TMI.2016.2538465

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

[40] Patel, H., & Zaveri, M. (2011). Credit card fraud detection using neural network. International Journal of Innovative Research in Computer and Communication Engineering, 1(2), 1–6. https://www.ijircce.com/upload/2011/october/1_Credit.pdf

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

[42] Alom, Md Zahangir, et al. "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches." ArXiv:1803.01164 [Cs], 12 Sept. 2018, arxiv.org/abs/1803.01164.

[43] Wang, Weibin, et al. "Medical Image Classification Using Deep Learning." Intelligent Systems Reference Library, 19 Nov. 2019, pp. 33–51, https://doi.org/10.1007/978-3-030-32606-7_3.

[44] Nabati, R., & Qi, H. (2019). "RRPN: Radar Region Proposal Network for Object Detection in Autonomous Vehicles." 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 3093-3097, doi: 10.1109/ICIP.2019.8803392.

[45] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[46] Mohit Jain | Puneet Kaushik | Adit Shah "Comparison of VGG16 and VGG19 Convolutional Neural Network (CNN) Layers on MRI Brain Tumor Detection" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-1 | Issue-1, December 2016, pp.275-280,
URL: https://www.ijtsrd.com/papers/ijtsrd3542.pdf

[47] Stupp, R., Taillibert, S., Kanner, A., Read, W., Steinberg, D. M., Lhermitte, B., Toms, S., Idbaih, A., Ahluwalia, M. S., Fink, K., Di Meco, F., Lieberman, F., Zhu, J.-J., Stragliotto, G., Tran, D. D., Brem, S., Hottinger, A., Kirson, E. D., Lavy-Shahaf, G., … Hegi, M. E. (2017). Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma: A randomized clinical trial. *JAMA, 318*(23), 2306–2316.
https://doi.org/10.1001/jama.2017.18718

[48] Raymaekers, J., Verbeke, W., & Verdonck, T. (2021). Weight-of-evidence 2.0 with shrinkage and spline-binning. arXiv preprint arXiv:2101.01494. Retrieved from https://arxiv.org/abs/2101.01494