



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IX **Month of publication:** September 2026

DOI: <https://doi.org/10.22214/ijraset.2026.84942>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Efficient Attention Pruning for Resource-Constrained Super-Resolution Using Statistical Thresholds and Quantum-Guided Sparsity

L. Nitheesh Kumar Reddy¹, Dr. K. Padma Priya²

¹M.Tech, ²Professor, Department of Electronics and Communication Engineering UCEK, JNTUK, Kakinada, Andhra Pradesh, India

Abstract—Non-local attention improves single-image super-resolution (SISR) by letting every query position aggregate information from all other positions, but the softmax normalisation at its core assigns strictly positive weight to every candidate regardless of relevance, diluting genuinely informative long-range correlations once the candidate set is large. This paper proposes Adaptive Sparse Attention (ASA), a non-local attention mechanism that replaces the fixed softmax support with a per-query threshold $\tau = \mu(s) + \alpha \cdot \sigma(s)$ derived from the mean and standard deviation of each query's own similarity-score distribution, so that only statistically above-average candidates are retained. This further proposes two extensions: SCASA-TT, which compresses the query/key/value projections with a Tensor-Train (TT) factorisation and adds a complementary spatial-coupling sparsity mask, and Quantum ASA, in which a 4-qubit variational quantum circuit (VQC) predicts the threshold coefficient α per query rather than using a fixed constant. All three variants, together with a non-attentive convolutional baseline, share a common backbone and are evaluated using Y-channel PSNR/SSIM on Set5, Set14, BSD100 and Urban100. Quantum ASA attains the best reconstruction quality of every method on every dataset and scale factor tested, with the largest gains observed on Urban100 and Manga109, the two benchmarks richest in repeated structure, consistent with the premise that non-local self-similarity is a more exploitable prior precisely where such redundancy is abundant. An analytical hardware-cost model projects a 73.7% reduction in multiply-accumulate operations and a 74.8% reduction in memory bandwidth for ASA-sparsified attention relative to dense attention at matched sequence length.

Keywords: Single image super-resolution, non-local attention, sparse attention, tensor-train decomposition, variational quantum circuit, resource-constrained deep learning.

I. INTRODUCTION

Non-local attention mechanisms exploit the self-similarity commonly present in natural images (repeated textures, structural edges, periodic patterns) by allowing every query position in a feature map to aggregate information from every other position, weighted by learned similarity. This has produced substantial gains in single-image super-resolution (SISR), where recovering fine detail from a low-resolution (LR) input benefits from borrowing evidence from distant but structurally similar regions of the same image [1], [2].

The mechanism has a structural weakness, however. Standard non-local attention computes a similarity score between a query and every candidate position and normalises the full set through softmax. Softmax assigns strictly positive weight to every candidate, irrespective of relevance: as the non-local sequence length N grows, softmax's output over N positions can approach a near-uniform distribution once genuine similarity is confined to a small subset, diluting the useful signal with vanishingly small but nonzero contributions from irrelevant positions. This dilution problem is compounded by a computational one: dense non-local attention over N positions incurs $O(N^2)$ similarity evaluation and weighted aggregation, a cost that is difficult to justify on resource-constrained hardware when most of that computation is, by the dilution argument above, spent on candidates that contribute little.

This paper addresses both problems jointly through a family of sparse attention mechanisms of increasing sophistication:

- Adaptive Sparse Attention (ASA), which replaces the fixed softmax support with a statistically-derived, per-query threshold, retaining only candidates whose similarity score exceeds their own query's mean plus a scalar multiple of its standard deviation, and discarding the remainder entirely rather than assigning them vanishing weight.

- SCASA-TT, a parameter-efficient extension that factorises the query/key/value projections with a Tensor-Train (TT) decomposition and unions the statistical mask with a second, spatial-coupling mask that additionally favours candidates near the query.
- Quantum ASA, an exploratory variant in which the scalar threshold coefficient is replaced by the output of a small variational quantum circuit (VQC), conditioned on the similarity-score statistics of the current query, allowing a learned, input-adaptive threshold rather than a single global constant.

All three variants, together with a non-attentive convolutional baseline for comparison, are evaluated on five standard SISR benchmarks (Set5, Set14, BSD100, Urban100, Manga109) at $\times 2$, $\times 3$ and $\times 4$ scale factors, together with an analytical hardware-cost model quantifying the multiply-accumulate, memory, bandwidth and energy reduction attributable to the proposed sparsification at matched sequence length.

II. RELATED WORK

Non-local neural networks were introduced by Wang *et al.* [1] as a differentiable generalisation of classical non-local means filtering for high-level vision tasks, and were subsequently adopted for SISR to exploit cross-scale and within-image self-similarity [2]. Mei *et al.* proposed non-local sparse attention (NLSN) [2], which uses locality-sensitive hashing to group similar positions before attention, reducing computation but retaining a learned, fixed hashing structure rather than a per-query statistical criterion. Xia *et al.* [3] proposed efficient non-local contrastive attention (ENLCN), which uses a contrastive learning objective to encourage the attention distribution to concentrate on genuinely similar positions. Most directly related to this work, Su *et al.* [4] identified the softmax dilution problem addressed here and proposed high-similarity-pass attention (HSPAN), which this work builds on and substantially reduces in scale and computational footprint, targeting resource-constrained deployment rather than maximum reconstruction quality. Separately, sparsemax [5] addressed a related dilution problem in classification and attention contexts by projecting onto the probability simplex rather than exponentiating, producing sparse but not adaptively-thresholded output. Tensor-Train (TT) decomposition [6] has been used to compress large weight tensors in neural networks with minimal accuracy loss [7], motivating its use here to offset the parameter cost of the additional query/key/value projections. Variational quantum circuits, implemented via automatic-differentiation frameworks such as PennyLane [8], allow small parametrised quantum circuits to be trained jointly with classical neural networks through ordinary backpropagation, which this work uses to predict an input-conditioned sparsification threshold.

A separate line of work targets parameter and computational efficiency directly within the accuracy-first CNN paradigm, without any non-local mechanism. IMDN [9] progressively splits and refines feature channels across a cascade of distillation blocks, retaining only a fraction of each block's output for further refinement while forwarding the remainder toward reconstruction. RFDN [10] refines this distillation mechanism further with a more parameter-efficient shallow residual block, winning the AIM 2020 efficient super-resolution challenge. Both achieve parameter counts in the 600,000-to-900,000 range, an order of magnitude smaller than EDSR [20] or RCAN [21] though still roughly 1.2–1.9 $\times$ larger than any variant developed here (483,587–496,107 parameters, Table I), while remaining purely convolutional and local in receptive field. This positions IMDN and RFDN as efficiency-oriented but non-local-blind: they address the parameter-cost side of the trade-off this work targets, but not the dilution and quadratic-cost problems of non-local attention motivating the ASA family, since they do not attempt non-local modelling at all.

III. PROPOSED METHOD

A. Backbone Architecture

All four model variants share a common convolutional backbone, AdaptiveSRNet: a 3×3 head convolution, four residual blocks (each comprising a Conv 3×3 , ReLU and Conv 3×3 layer, with a residual skip) split evenly around a single attention module, a fusion convolution with a long residual skip from the head, and a PixelShuffle $\times s$ upsampling layer for scale factor $s \in \{2, 3, 4\}$. The Baseline variant replaces the attention module with an identity mapping, isolating the accuracy and cost contribution attributable to non-local modelling. All four variants use $C = 64$ channels, giving near-identical parameter budgets (Table I) so that differences in reconstruction quality can be attributed to the attention mechanism rather than to incidental capacity differences.

B. Adaptive Sparse Attention (ASA)

Given an input feature map, three 1×1 convolutions produce query, key and value projections Q, K, V . For each query position, the similarity vector against all N candidate positions is computed as $s = Q^T K / \sqrt{C}$, exactly as in standard non-local attention. Rather than passing s directly through softmax, ASA first computes an adaptive threshold from the statistics of s itself:

$$\tau = \mu(s) + \alpha \cdot \sigma(s) \quad (1)$$

where $\mu(s)$ and $\sigma(s)$ are the mean and standard deviation of the similarity scores for that query, and α is a scalar hyperparameter controlling sparsity aggressiveness ($\alpha = 0.5$ in the fixed-threshold configuration evaluated here). All entries of s below τ are masked to zero; the surviving entries are passed through a softmax restricted to the retained support, so the final attention weights over the active set still sum to one. Because τ is derived independently for every query position from that query's own similarity statistics, the resulting sparsity pattern is fully adaptive: queries with a small number of strongly similar candidates retain a small active set, while queries with more broadly distributed similarity retain a larger one. Empirically, this mechanism prunes approximately 79% of candidate positions on average while improving reconstruction quality over the Baseline, consistent with the dilution argument of Section II: the pruned positions were contributing noise rather than useful signal under the dense softmax formulation. Fig. 1 shows the module's position in the full pipeline and its internal structure.

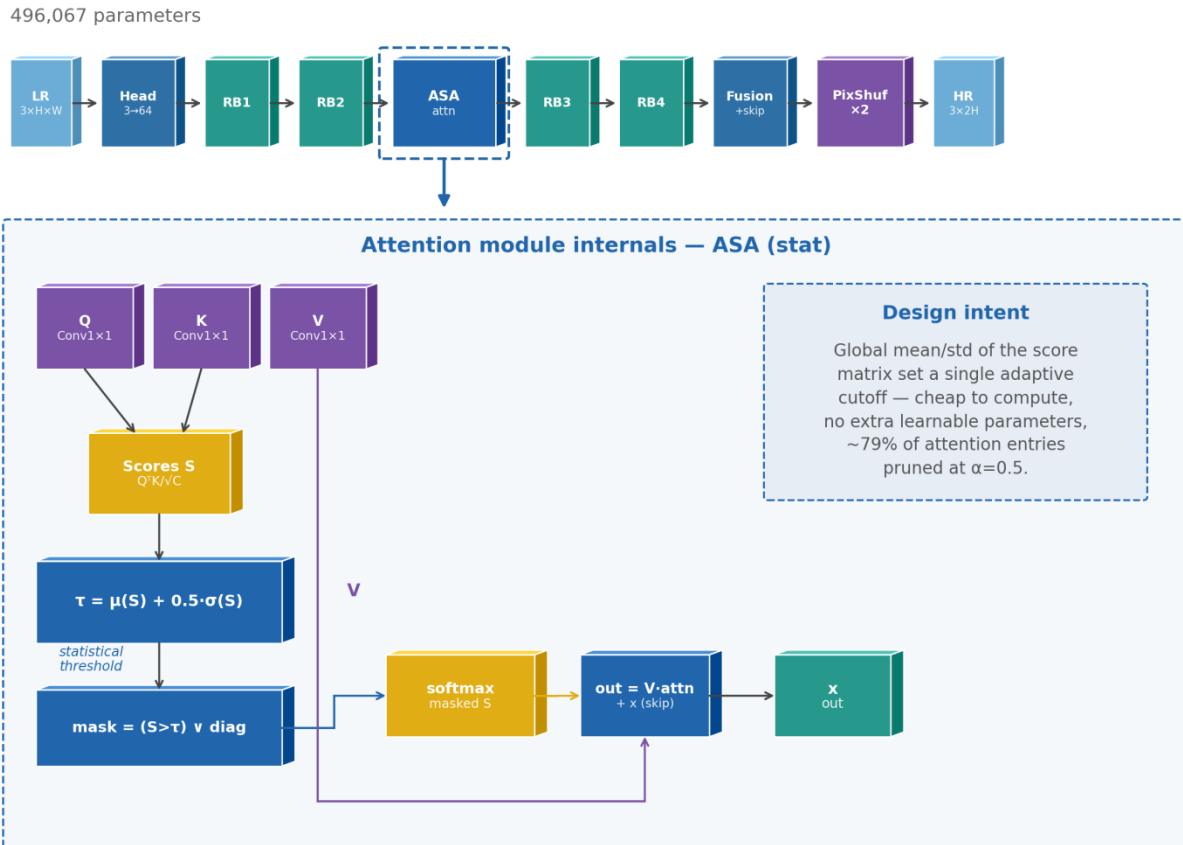


Fig. 1. ASA (stat): position in the AdaptiveSRNet pipeline (top strip) and attention-module internals (bottom)

C. SCASA-TT: Spatial-Coupled ASA with Tensor-Train Compression

SCASA-TT extends ASA along two complementary directions. First, the Q/K/V projection convolutions are replaced with Tensor-Train-compressed projections (factor shape 8×8 , TT-rank 4), substantially reducing the parameter count of the projection layers relative to a standard 1×1 convolution of the same channel width. Second, SCASA-TT computes two independent sparsity masks and takes their union before the final softmax: the statistical mask of Eq. (1), and a second, spatial-coupling mask computed via a small auxiliary relaxation network that additionally down-weights candidates based on their spatial displacement from the query position.

A candidate position is retained if selected by either criterion: genuinely high feature similarity, or a combination of moderate similarity and spatial proximity, reflecting the observation that self-similar structure in natural images is often, though not always, spatially clustered. Fig. 2 shows the module's internal structure, including the two parallel masking branches.

485,315 parameters

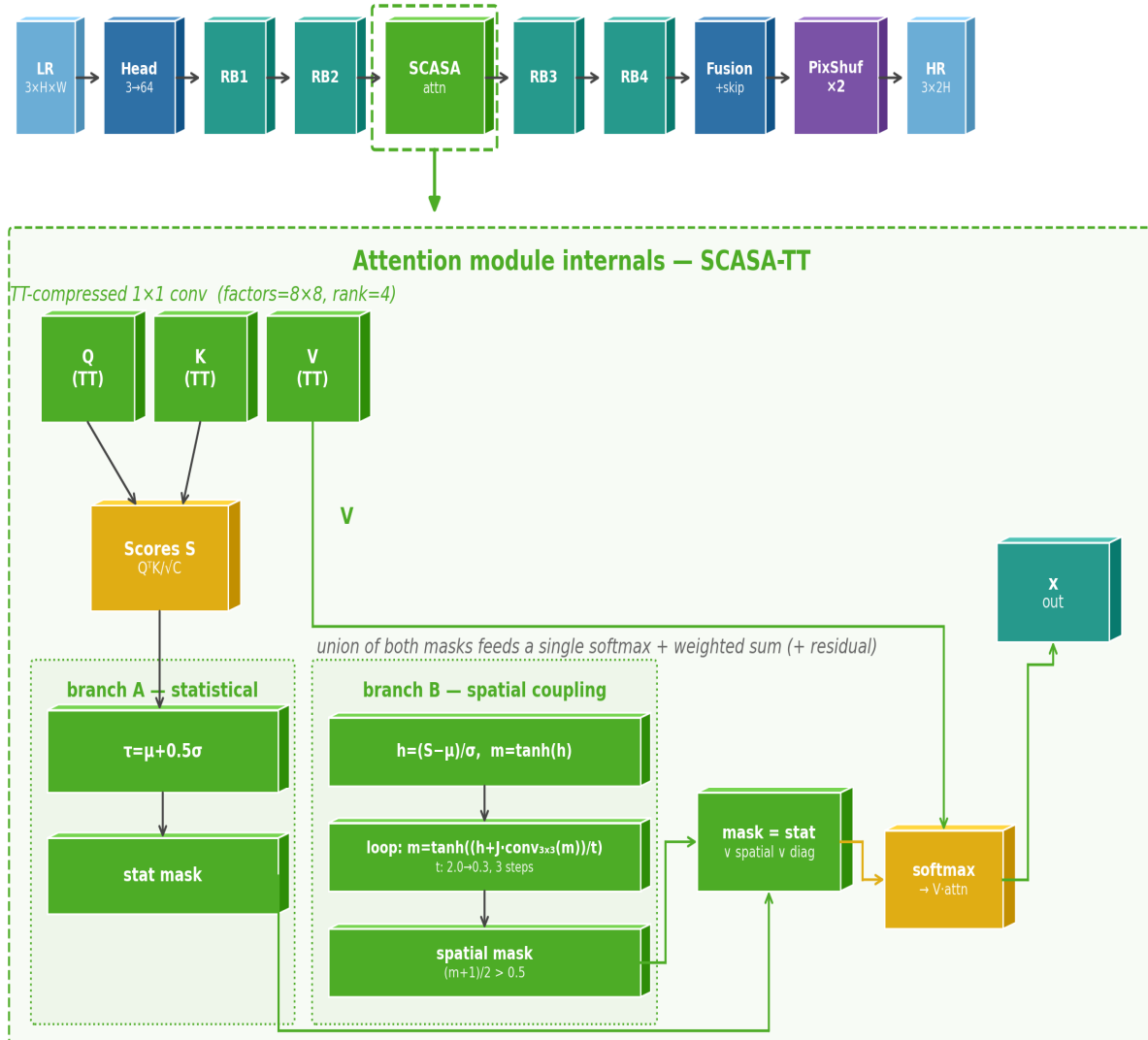


Fig. 2. SCASA-TT: position in the pipeline and attention-module internals, the Tensor-Train-compressed Q/K/V projections and the two parallel masking branches

D. Quantum ASA

Quantum ASA retains the overall ASA structure but replaces the fixed scalar α with a value predicted by a 4-qubit variational quantum circuit, conditioned on four summary statistics of the current query's similarity scores (mean, standard deviation, skewness and kurtosis). These statistics are encoded into the initial qubit rotation angles via angle embedding; three subsequent layers each apply a learnable $\text{Rot}(R_z, R_y, R_z)$ gate to every qubit followed by a ring of CNOT entangling gates (36 trainable parameters in total). The circuit output is obtained by measuring the Pauli-Z expectation value of one qubit and mapping it, via a fixed affine transformation, into the valid threshold range $\alpha \in [0.1, 1.5]$. Because the circuit is implemented in PennyLane [8], which supports automatic differentiation through quantum circuit simulation, the rotation angles are trained jointly with the rest of the network via ordinary backpropagation, with no separate optimisation procedure required. Across a 200-epoch training run, the circuit-predicted threshold converges to $\alpha \approx 0.4356$, marginally below the fixed statistical value of 0.5, corresponding to a somewhat more permissive (less aggressive) learned cutoff than the hand-derived rule. Fig. 3 shows the module's internal structure and the position of the VQC within it.

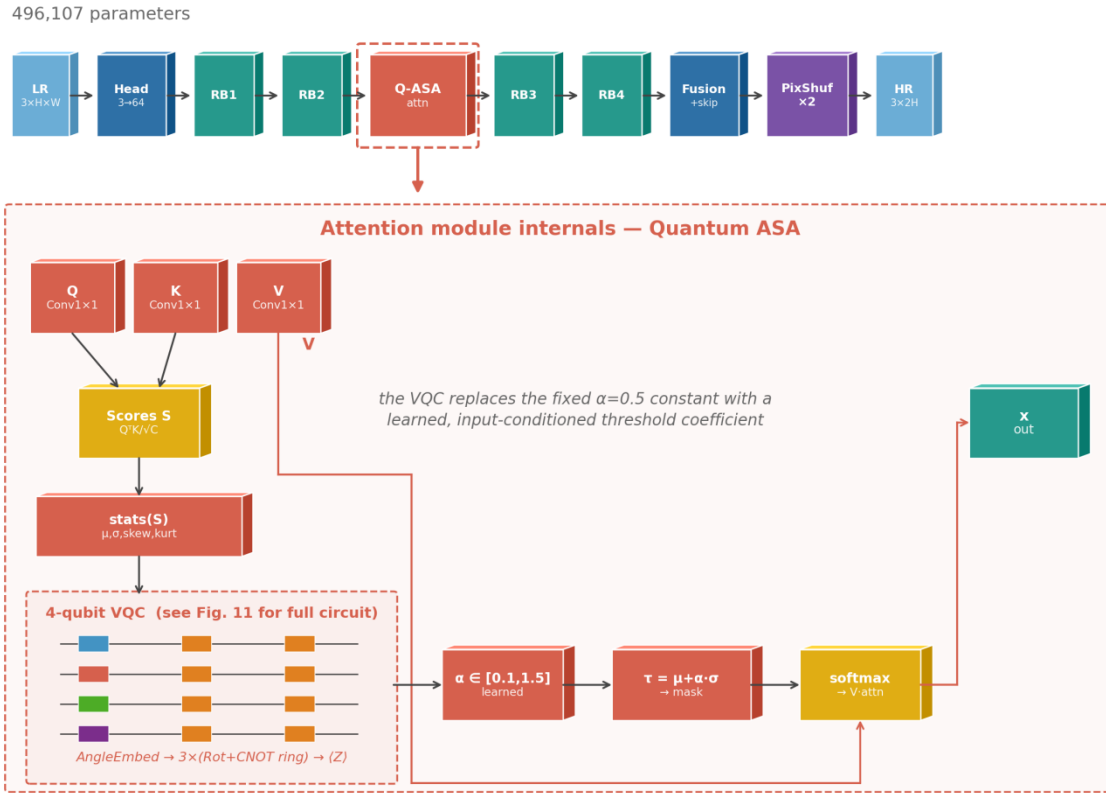


Fig. 3. Quantum ASA: position in the pipeline and attention-module internals, the 4-qubit VQC that consumes the query's similarity statistics and outputs the gating parameter α in place of the fixed constant.

TABLE I

PARAMETER COUNT OF THE FOUR MODEL VARIANTS (C = 64)

Variant	Attention Mechanism	Parameters
Baseline	None (identity)	483,587
ASA (stat)	$\tau = \mu + \alpha \sigma$	496,067
SCASA-TT	Stat. + spatial-coupling mask, TT proj.	485,315
Quantum ASA	4-qubit VQC-predicted threshold	496,107

IV. EXPERIMENTAL SETUP

A. Datasets

Training data is drawn from DIV2K [13]; 800 images are sampled (fixed seed 42) as the training pool for every variant and scale factor. A 48×48 LR patch is randomly cropped per image, with the matching HR patch extracted at the scale-multiplied location, and augmented on-the-fly with random horizontal flip, vertical flip, and 90°-multiple rotation (each applied independently with probability 0.5). LR images are pre-generated via bicubic downsampling rather than downsampled on-the-fly during training.

Evaluation uses five standard benchmarks: Set5 [15], Set14 [16], BSD100 [17], Urban100 [18], and Manga109 [19]. At ×2, Set5, Set14, BSD100 and Urban100 use standard pre-packaged LR/HR pairs. At ×3 and ×4, evaluation covers Set5, Set14, Urban100 and Manga109 (BSD100 is not included at these scales); Manga109 provides raw page scans rather than pre-packaged folders, and the Urban100 distribution used lacks a pre-packaged ×3 folder, so in both cases the LR/HR pair is synthesised via mod-crop and bicubic downsampling. Urban100 and Manga109 were included deliberately: since this work's central premise is that non-local self-similarity is a genuinely exploitable prior, these are the two benchmarks most likely to reveal a measurable benefit, being constructed around (Urban100) and dominated by (Manga109) abundant repeated structure.

B. Evaluation Metrics

Reconstruction quality is measured via PSNR and SSIM [14], computed on the Y (luminance) channel of YCbCr following standard SISR convention, with RGB-to-Y conversion using ITU-R BT.601 coefficients. A 4-pixel border is discarded from each edge before evaluation, since border pixels are disproportionately affected by patch-based-inference boundary artefacts rather than genuine reconstruction error. PSNR/SSIM are computed using scikit-image's implementations with data_range = 1.0, and reported values are means across every image in a test dataset.

For images I and G of height H and width W , mean squared error is $MSE(I, G) = (1/HW) \sum_i \sum_j [I(i,j) - G(i,j)]^2$, and $PSNR(I, G) = 10 \log_{10} (MAX^2/MSE(I, G))$ with $MAX = 255$. PSNR is unbounded and penalises every pixel-wise deviation equally regardless of perceptual salience. SSIM instead compares local luminance, contrast and structure statistics over an 11×11 Gaussian window (standard deviation 1.5 px), using stabilising constants $C_1 = (K_1 \cdot MAX)^2$ and $C_2 = (K_2 \cdot MAX)^2$ with $K_1 = 0.01$, $K_2 = 0.03$. SSIM is comparatively more sensitive to structural distortion blurring, ringing, misaligned texture than to a uniform intensity offset, a property this work exploits throughout to distinguish indiscriminate smoothing from genuine structural fidelity.

C. Implementation and Training Details

All models are implemented in PyTorch; Quantum ASA additionally uses PennyLane with the lightning.qubit simulator backend (falling back to default.qubit) and adjoint differentiation for circuit gradients. All variants train with the Adam optimiser and L1 reconstruction loss for 200 epochs, with an initial learning rate of 1×10^{-4} decayed by a factor of 0.5 at epochs 100, 150 and 180. Baseline, ASA (stat) and SCASA-TT use a batch size of 16; Quantum ASA uses a reduced batch size of 8 to accommodate the memory cost of quantum-circuit simulation. For Quantum ASA, the circuit parameters use a separate learning rate of 1×10^{-4} (increased from an initial 5×10^{-5}), a sparsity-regularisation loss term with weight 0.5, and a target active ratio of 0.20. Each scale factor ($\times 2$, $\times 3$, $\times 4$) is trained independently from random initialisation; the $\times 2$ configuration was trained for all four variants, while at $\times 3/\times 4$ only Baseline and Quantum ASA were trained, representing the two extremes of the design space within the available compute budget.

Inference on full-resolution images uses patch-wise sliding-window processing with 48×48 LR patches and 16-pixel overlap (stride 32). Each patch is upsampled independently, and overlapping output regions are blended using a 2-D Hann window, so each patch's contribution is smoothly down-weighted toward its edges and full-weighted at its centre; the accumulated canvas is normalised by accumulated weight at every pixel. This removes visible patch-boundary seams and is applied identically across all variants and scale factors. Training was carried out on Kaggle notebook sessions with GPU acceleration; evaluation, qualitative visualisation, and efficiency analysis were subsequently carried out on Google Colab using the trained checkpoints.

D. Efficiency Measurement Protocol

Parameter count is computed by summing elements across all trainable parameters. MACs and derived FLOPs are measured using the thop profiling library on a single $48 \times 48 \times 3$ input patch. Wall-clock inference time is measured by timing five repeated sliding-window inference calls on a representative Set5 image, discarding the first (warm-up) call and averaging the remainder, with CUDA synchronisation enforced immediately before and after each timed call. Two limitations of this protocol are worth noting: thop counts MACs by tracing standard layer types, so operations outside this set including ASA's masking/thresholding and Quantum ASA's circuit evaluation are not automatically counted; and measuring latency on a single representative image trades statistical robustness for controlled cross-variant comparison. The separately-computed analytical hardware-cost model (Section V-E) addresses the first limitation by estimating the sparsification-specific cost directly, at a fixed feature-map size ($C = 64$, $H = W = 48$) and the measured 79% sparsity level, using a fixed per-MAC energy cost of 0.1 pJ and an assumed peak throughput of 1 TFLOPS.

V. RESULTS AND DISCUSSION

A. Quantitative Results at $\times 2$

Table II reports Y-channel PSNR/SSIM at $\times 2$ across all four datasets. Quantum ASA attains the best PSNR and SSIM of every method on every dataset, with the margin over Baseline ranging from +0.24 dB (BSD100) to +0.56 dB (Urban100); the largest gain is observed on Urban100, the benchmark explicitly constructed to contain abundant repeated architectural structure. ASA (stat) and SCASA-TT improve over Baseline on every dataset as well, though by a substantially smaller margin (+0.01 to +0.07 dB), and are essentially indistinguishable from one another in accuracy despite SCASA-TT adding only 1,728 parameters over Baseline (a 0.36% increase) compared with ASA (stat)'s 12,480-parameter (2.58%) overhead, indicating that the Tensor-Train-compressed projections

in SCASA-TT retain essentially all of the representational benefit of the full-rank ASA (stat) projections at a fraction of the parameter cost. Fig. 4 presents the same comparison graphically across all four datasets.

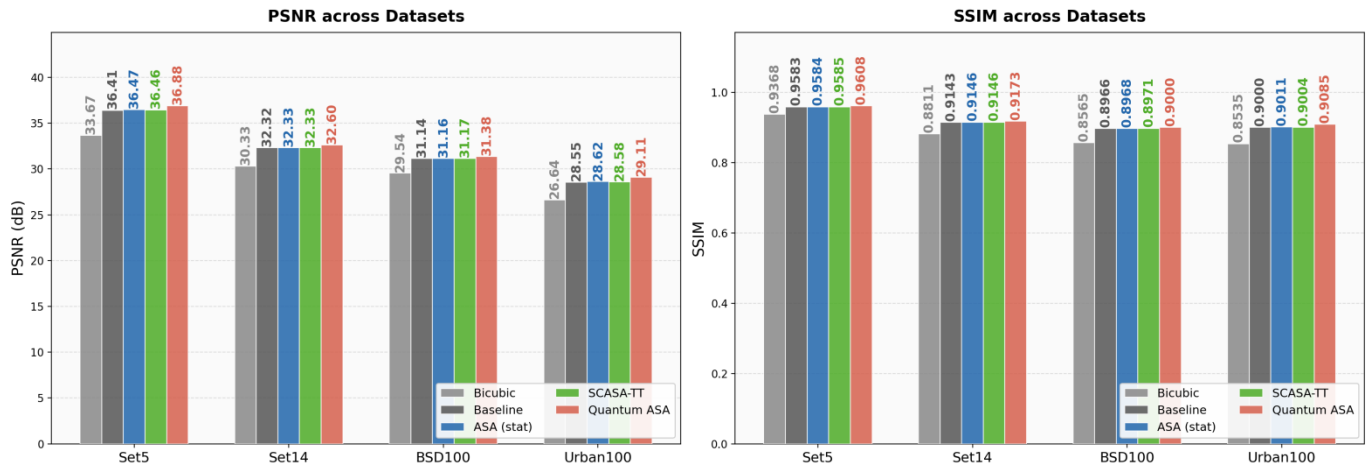


Fig. 4. Y-channel PSNR and SSIM across Set5, Set14, BSD100 and Urban100 at $\times 2$

TABLE II

Y-CHANNEL PSNR (DB) / SSIM AT $\times 2$

Method	Set5	Set14	BSD100	Urban100
Bicubic	33.67 / .9368	30.33 / .8811	29.54 / .8565	26.64 / .8535
Baseline	36.41 / .9583	32.32 / .9143	31.14 / .8966	28.55 / .9000
ASA (stat)	36.47 / .9584	32.33 / .9146	31.16 / .8968	28.62 / .9011
SCASA-TT	36.46 / .9585	32.33 / .9146	31.17 / .8971	28.58 / .9004
Quantum ASA	36.88 / .9608	32.60 / .9173	31.38 / .9000	29.11 / .9085

B. Generalisation to $\times 3$ and $\times 4$

Table III reports the $\times 3$ and $\times 4$ results for Baseline and Quantum ASA. The Quantum ASA margin over Baseline widens at higher scale factors, and is widest of all on Manga109 at $\times 3$ (+1.24 dB), a dataset dominated by printed text and line art. This degree of exact, repeated local structure, present even under heavy downsampling, plausibly explains why the non-local self-similarity mechanism extracts a disproportionately large benefit on this dataset specifically.

TABLE III

Y-CHANNEL PSNR (DB) / SSIM, BASELINE VS. QUANTUM ASA

Scale	Dataset	Baseline	Quantum ASA
$\times 3$	Set5	32.29 / .9133	32.83 / .9209
	Set14	29.04 / .8340	29.34 / .8399
	Urban100	25.17 / .7914	25.61 / .8067
	Manga109	29.33 / .9056	30.57 / .9210
$\times 4$	Set5	30.18 / .8680	30.51 / .8762
	Set14	27.29 / .7672	27.52 / .7740
	Urban100	24.16 / .7270	24.43 / .7389
	Manga109	26.79 / .8472	27.41 / .8621

C. Qualitative Results

Fig. 5 shows zoomed reconstructions on two Urban100 images at $\times 2$, with whole-image Y-channel PSNR/SSIM captioned above each crop. This comparison surfaces an instructive dissociation between PSNR and SSIM: on *img_004*, Bicubic interpolation attains the highest raw PSNR of any method (22.45 dB) despite the lowest SSIM (0.8643), a known artefact of PSNR on highly repetitive high-frequency textures, where over-smoothing reduces squared pixel error even as it destroys structural detail, which is why this work reports both metrics throughout rather than PSNR alone.

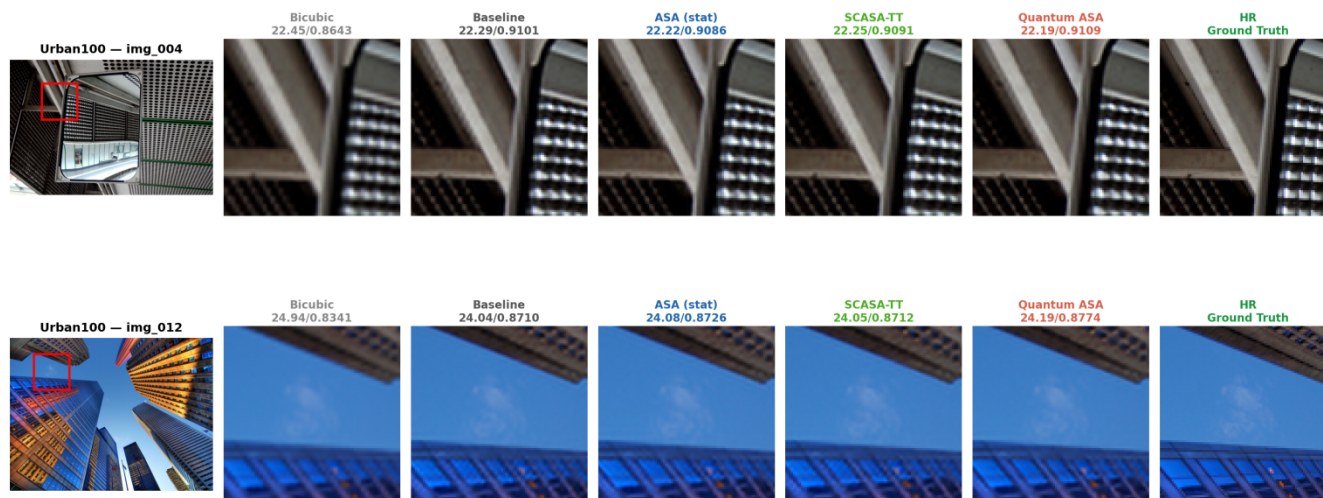


Fig. 5. Qualitative comparison on two Urban100 images at $\times 2$

D. Attention-Map Interpretability

To visualise what the sparsification mechanism is actually discarding, Fig. 6 compares the non-local attention weight assigned to a fixed query position by (i) a generic dense-attention reference computed directly from raw features with no thresholding, (ii) the Baseline model's own dense similarity map, and (iii)–(v) the sparsified maps produced by ASA (stat), SCASA-TT and Quantum ASA. Both dense references show visibly diffuse, near-uniform-looking support across almost the entire candidate set regardless of whether the underlying features came from a model trained with attention: a direct, model-agnostic visual confirmation of the softmax-dilution phenomenon motivating this work. The three sparsified variants, in contrast, retain a visually coherent, image-dependent subset of positions concentrated along repeated edges, textures or structural lines that a human observer would judge as genuinely self-similar to the query patch.

Fig. 7 shows the training dynamics underlying this behaviour across the 200-epoch $\times 2$ run: L1 training loss, the attention active ratio (one minus the pruned fraction), and, for Quantum ASA, the evolution of the circuit-predicted gating parameter α . The active ratio reveals a finding that is, at first glance, counter-intuitive: SCASA-TT converges to the highest active ratio of the three sparsified variants (≈ 0.680 , i.e. approximately 32% of connections pruned), Quantum ASA converges to an intermediate active ratio (≈ 0.244 , $\approx 76\%$ pruned), and ASA (stat), the variant using the simplest, fixed statistical threshold, converges to the lowest active ratio of the three (≈ 0.153 , $\approx 85\%$ pruned), making it the most aggressive pruner despite Table II showing it is not the best-performing sparse variant. This indicates that reconstruction quality in this family of models is governed more by which non-local connections are retained, through spatial coupling in SCASA-TT's case or a circuit-conditioned, per-query adaptive threshold in Quantum ASA's case, than by how aggressively the candidate set is pruned in absolute terms.

E. Resource Efficiency

Table IV reports measured parameter overhead and wall-clock sliding-window inference time. The parameter overhead of every sparsified variant remains under 2.6% relative to Baseline, but measured wall-clock inference time increases by a factor of roughly 6 to 9 depending on the variant. the present PyTorch/PennyLane implementation computes the full dense similarity matrix and, for Quantum ASA, evaluates the variational circuit and its automatic-differentiation graph, before any thresholding is applied, so the wall-clock cost reflects unexploited dense computation rather than the cost of the sparsified representation itself.

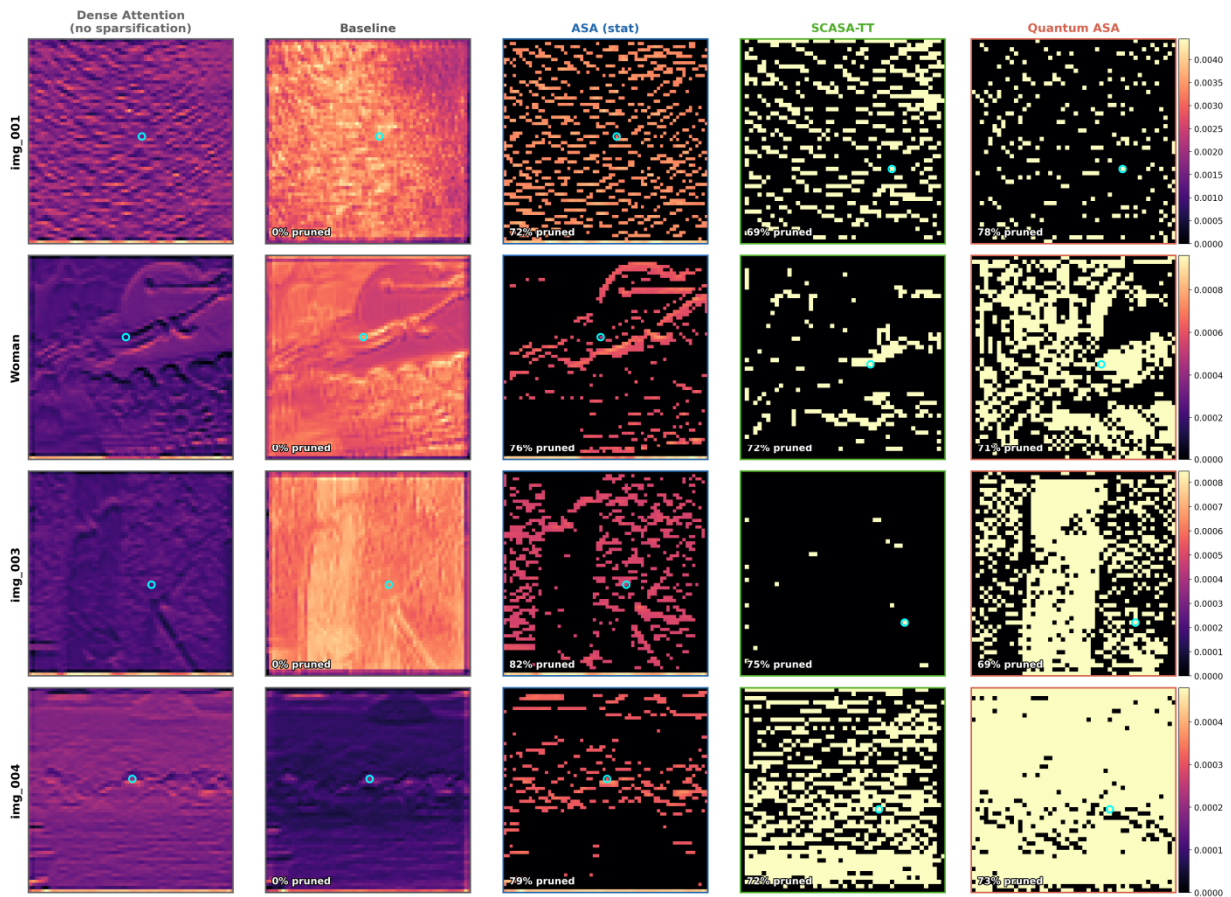


Fig. 6. Dense-versus-sparse attention maps for a representative query patch (cyan circle), across a generic dense reference, the Baseline model's own dense similarity map, and the three sparsified variants (ASA (stat), SCASA-TT, Quantum ASA). Brighter pixels indicate higher attention weight; the percentage of pruned (zero-weight) connections is annotated on each sparsified panel.

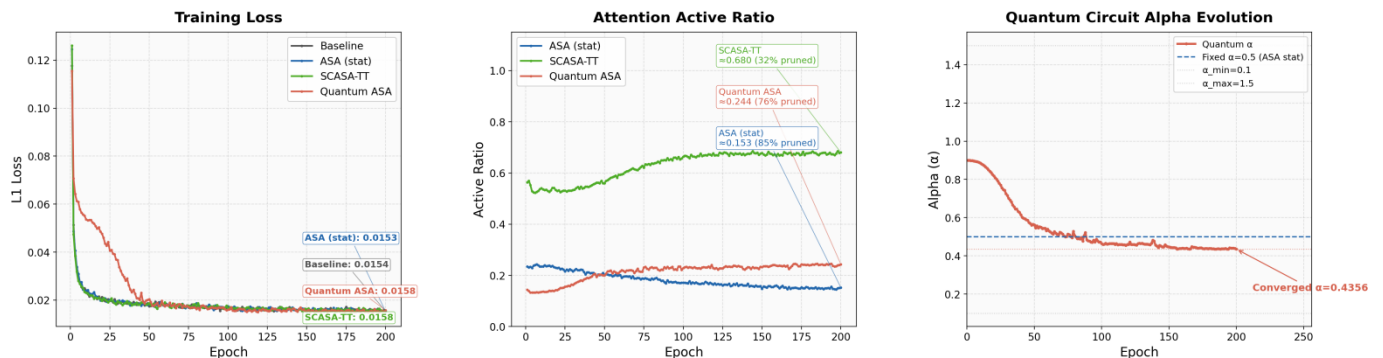


Fig. 7. Training dynamics across the four $\times 2$ model variants: L1 training loss (left), attention active ratio (centre), and Quantum ASA's circuit-predicted gating parameter α over training (right).

The separately-computed analytical hardware-cost model projects a 73.7% reduction in MACs, a 74.8% reduction in memory bandwidth, and a 73.7% reduction in energy for ASA-sparsified attention relative to dense attention at matched sequence length and the measured 79% sparsity level; realising this projection in wall-clock terms requires sparse-aware computation, such as specialised sparse-matrix kernels or, ultimately, dedicated hardware, which the present software-level implementation does not attempt. Fig. 8 presents this projection graphically, together with a sweep of projected MAC reduction against the threshold coefficient α .

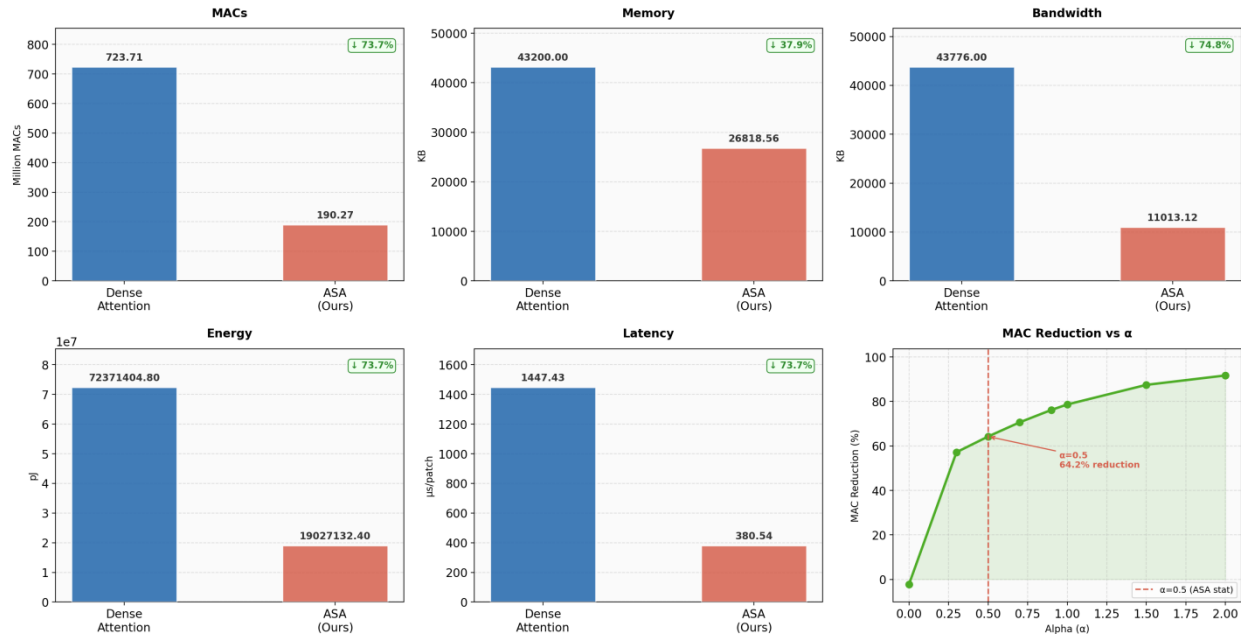


Fig. 8. Analytical hardware-cost comparison of dense non-local attention versus ASA-sparsified attention at matched sequence length ($C = 64$, $H = W = 48$, measured sparsity 79%): multiply-accumulate operations, memory footprint, memory bandwidth, estimated energy and estimated latency, together with a sweep of projected MAC reduction against α .

For additional context, on-disk model size at fp32 precision also compares favourably against several representative literature SISR models: published configurations of NLSN [2], ENLCN [3], DRLN [11] and the base HSPAN [4] architecture occupy roughly 130–180 MB, and SwinIR [12] (classical configuration) is 47.2 MB, versus under 2.1 MB for every variant developed here. This is noted only as broader size context, not a controlled comparison, since these models were trained and evaluated under different backbones, datasets and protocols; HSPAN’s larger footprint in particular reflects its published 192-channel, ten-module configuration against the 64 channels and single module used here, and Quantum ASA’s absolute PSNR (Table II) still trails well behind large SOTA models in the wider literature.

TABLE IV
PARAMETER COUNT AND WALL-CLOCK INFERENCE TIME

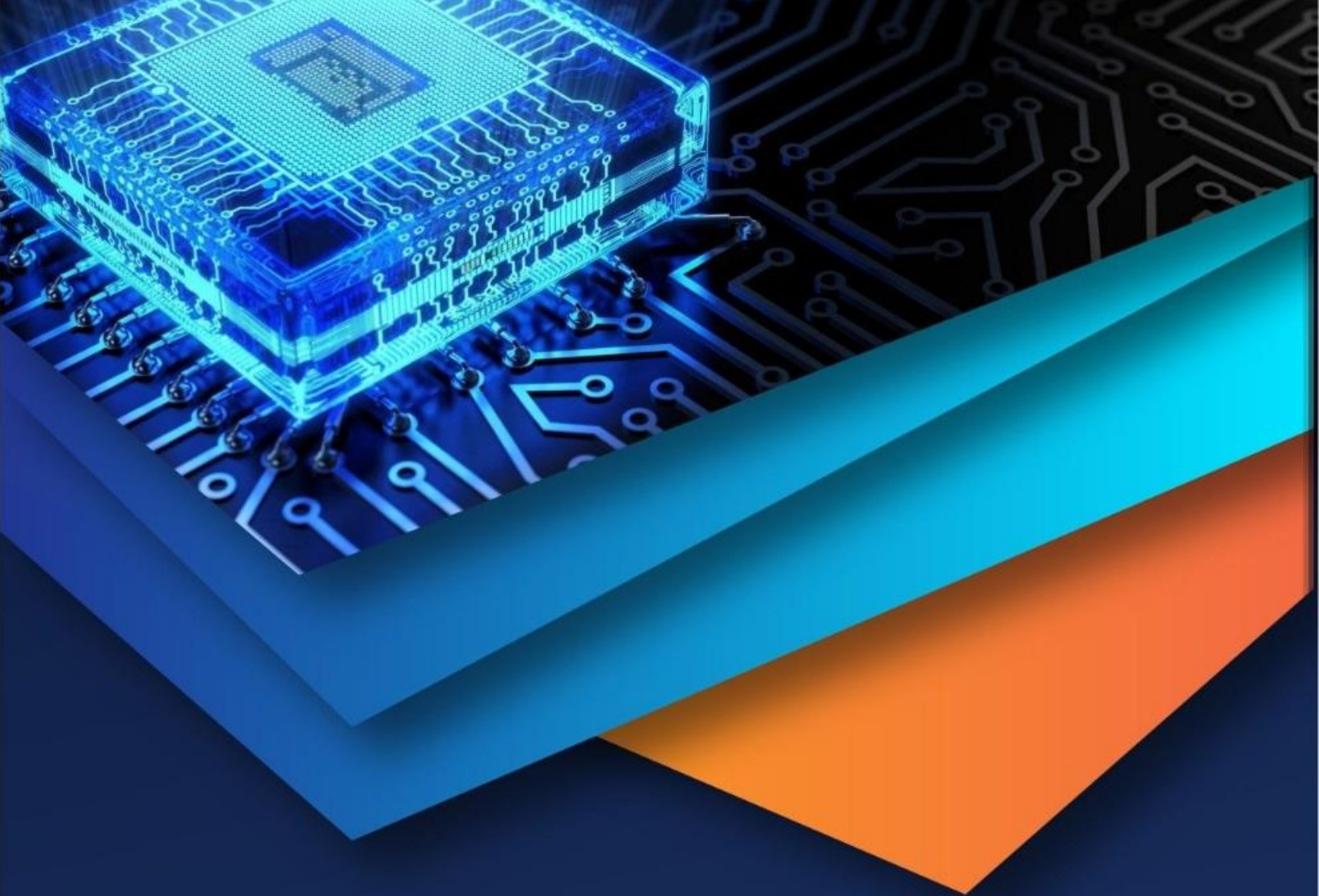
Variant	Params	Δ Params	Inference (ms)	Δ Inference (ms)
Baseline	483,587	–	82.9	–
ASA (stat)	496,067	+2.58%	530.5	+539.9%
SCASA-TT	485,315	+0.36%	641.8	+674.2%
Quantum ASA	496,107	+2.59%	753.1	+808.4%

VI. CONCLUSION

This paper proposed Adaptive Sparse Attention, a statistically-thresholded alternative to dense softmax non-local attention for SISR, together with a Tensor-Train-compressed and spatially-coupled extension (SCASA-TT) and a quantum-circuit-guided variant (Quantum ASA) in which the sparsification threshold is predicted rather than fixed. Across five standard SISR benchmarks and three scale factors, Quantum ASA consistently outperforms a matched-capacity non-attentive baseline and a fixed-threshold statistical variant, with the largest gains on datasets richest in repeated structure. An analytical hardware-cost model indicates substantial multiply-accumulate, bandwidth and energy reduction is achievable from the sparsification pattern learned by these methods; realising this reduction in wall-clock terms through sparse-aware computation is identified as the primary direction for future work.

REFERENCES

- [1] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7794–7803.
- [2] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 3516–3525.
- [3] B. Xia, Y. Hang, Y. Tian, W. Yang, Q. Liao, and J. Zhou, "Efficient non-local contrastive attention for image super-resolution," in Proc. AAAI Conf. Artif. Intell., vol. 36, no. 3, 2022, pp. 2759–2767.
- [4] J.-N. Su, M. Gan, G.-Y. Chen, W. Guo, and C. L. P. Chen, "High-similarity-pass attention for single image super-resolution," IEEE Trans. Image Process., vol. 33, pp. 610–624, 2024.
- [5] A. Martins and R. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in Proc. Int. Conf. Mach. Learn. (ICML), 2016, pp. 1614–1623.
- [6] I. V. Oseledets, "Tensor-train decomposition," SIAM J. Sci. Comput., vol. 33, no. 5, pp. 2295–2317, 2011.
- [7] A. Novikov, D. Podoprikhin, A. Osokin, and D. Vetrov, "Tensorizing neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2015, pp. 442–450.
- [8] V. Bergholm et al., "PennyLane: Automatic differentiation of hybrid quantum-classical computations," arXiv:1811.04968, 2018.
- [9] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 2024–2032.
- [10] J. Liu, J. Tang, and G. Wu, "Residual feature distillation network for lightweight image super-resolution," in Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW), 2020, pp. 41–55.
- [11] S. Anwar and N. Barnes, "Densely residual Laplacian super-resolution," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, pp. 1192–1204, 2022.
- [12] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin Transformer," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), 2021, pp. 1833–1844.
- [13] R. Timofte et al., "NTIRE 2017 challenge on single image super-resolution: Methods and results," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 1110–1121.
- [14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, 2004.
- [15] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in Proc. Brit. Mach. Vis. Conf. (BMVC), 2012, pp. 1–10.
- [16] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in Proc. Int. Conf. Curves Surf., 2010, pp. 711–730.
- [17] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV), 2001, pp. 416–423.
- [18] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 5197–5206.
- [19] Y. Matsui et al., "Sketch-based manga retrieval using Manga109 dataset," Multimedia Tools Appl., vol. 76, no. 20, pp. 21811–21838, 2017.
- [20] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 1132–1140.
- [21] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 286–301.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)