American Journal of Artificial Intelligence and Neural Networks

australiansciencejournals.com/ajainn

E-ISSN: 2688-1950

VOL 07 ISSUE 01 2026

Computationally Efficient CT Reconstruction via Structured Sparse Networks

Jeroen van Dijk[1], Anneke Smit[2], Thomas de Vries[3]*

Department of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands

***Corresponding author:** t.devries@tudelft.nl*

Abstract

Clinical deployment of learned CT reconstruction models requires attention to runtime, memory footprint, and throughput. Building on convolution–self-attention architectures such as CTLformer, this paper introduces lightweight convolution–attention blocks with structured sparsity and low-rank attention approximation. The design reduces quadratic attention costs while keeping long-range dependency modeling. Experiments are conducted on two CT reconstruction datasets totaling 40,000 slices. Compared with transformer-heavy reconstruction models and standard hybrid baselines, the proposed approach reduces GPU memory usage by 28%–40% and improves inference throughput by 1.6×–2.1×, while keeping reconstruction accuracy within 0.2–0.4 dB PSNR of the full model.

Keywords: *CT reconstruction; model efficiency; lightweight attention; structured sparsity; deployment*

1.INTRODUCTION

Deep learning–based methods have been widely adopted for CT reconstruction, particularly in low-dose and sparse-view settings where analytic reconstruction and classical regularization approaches often leave residual noise and streak artifacts [1,2]. Convolutional neural networks provide stable performance and efficient local feature extraction, but their limited receptive field constrains the suppression of structured artifacts that extend over long spatial ranges. To overcome this limitation, recent hybrid architectures that integrate convolutional layers with attention mechanisms have been proposed, showing improved detail preservation and artifact reduction compared with purely convolutional models, especially in regions affected by directional

streaks and global inconsistencies [3,4]. A representative convolution–attention reconstruction framework incorporates self-attention modules within a lightweight convolutional backbone to enhance low-dose CT reconstruction quality while maintaining spatial coherence [5]. These results demonstrate the potential of hybrid designs to balance local detail modeling and global context aggregation. Despite these advances in reconstruction quality, many existing studies primarily emphasize accuracy under controlled experimental conditions and provide limited analysis of computational efficiency, which remains a critical requirement for clinical deployment [6]. In routine clinical environments, CT reconstruction systems must satisfy strict constraints on inference latency, memory usage, and throughput. Reconstruction models are expected to process large image volumes continuously on shared hardware while maintaining stable performance across workloads and patient cohorts [7]. Attention-based architectures pose a particular challenge in this context, as standard self-attention mechanisms exhibit quadratic complexity with respect to feature size, resulting in high memory consumption and limited scalability for high-resolution CT images [8]. Consequently, models that achieve strong offline reconstruction performance may be difficult to integrate into clinical pipelines with fixed hardware budgets. To mitigate the computational burden of attention mechanisms, several efficiency-oriented strategies have been explored in the broader deep learning literature. Low-rank and kernel-based attention approximations replace dense attention computation with more efficient formulations, significantly reducing memory and computational cost while retaining the ability to model long-range dependencies [9,10]. Structured sparsity and pruning methods aim to remove redundant computation by enforcing sparsity patterns within network weights or activations, enabling more efficient execution on modern hardware [11,12]. Recent findings indicate that sparsity patterns aligned with hardware constraints can yield practical reductions in inference time and memory usage when incorporated during training rather than applied as post hoc pruning [13,14]. However, these efficiency-oriented attention strategies have rarely been systematically integrated into CT reconstruction models. Within CT reconstruction research, efficiency is often treated as a secondary consideration. Many studies focus on image quality improvements over iterative or unrolled reconstruction baselines, while reporting limited information on peak memory consumption, runtime, or scalability across input resolutions [15]. Even hybrid convolution–attention models that effectively suppress artifacts may rely on dense attention blocks that dominate overall computational cost [16,17]. In addition, experimental evaluation is frequently conducted on relatively small test sets or simplified

settings, which limits assessment of whether reported efficiency gains remain stable under realistic throughput and deployment conditions [18]. As a result, the practical feasibility of attention-enhanced reconstruction models for routine clinical use remains uncertain. These observations highlight a gap between reconstruction quality improvements and deployable efficiency in attention-based CT reconstruction. Bridging this gap requires architectural designs that explicitly target the dominant sources of computational cost while preserving the long-range dependency modeling that contributes to artifact suppression. Lightweight attention formulations, structured sparsity, and cost-aware design principles provide promising directions, but their integration within reconstruction-specific architectures has not been fully explored.

In this work, an efficient CT reconstruction framework is presented based on lightweight convolution–attention blocks with structured sparsity and low-rank attention approximation. The proposed design directly addresses the primary computational bottlenecks of attention-based reconstruction while maintaining the capacity to model global contextual information relevant to structured artifact suppression. Instead of relying on transformer-heavy backbones, the framework introduces efficiency-oriented building blocks that reduce memory usage and improve inference throughput. The method is evaluated on two CT reconstruction datasets comprising 40,000 slices. Experimental results demonstrate a reduction in GPU memory usage of 28%–40% and an increase in inference throughput by 1.6×–2.1×, while maintaining reconstruction accuracy within 0.2–0.4 dB PSNR of the full model. These findings indicate that convolution–attention reconstruction can be adapted to clinical hardware constraints without substantial loss of image quality, supporting more practical deployment in routine CT workflows.

2.Materials and Methods

2.1 Sample Description and Imaging Conditions

This study used two CT reconstruction datasets collected under routine clinical protocols. The combined dataset included 40,000 axial slices from adult chest and abdominal examinations. Images were reconstructed from low-dose and standard-dose acquisitions using consistent scanner geometry and reconstruction kernels. The in-plane resolution ranged from 0.6 to 0.9 mm, and slice thickness varied between 1.0 and 2.5 mm. The data contain diverse anatomical structures and structured noise patterns commonly observed in clinical CT. All images were anonymized prior to analysis.

2.2 Experimental Design and Reference Methods

A comparative experimental design was used to evaluate reconstruction quality and efficiency. The experimental group applied the proposed lightweight convolution–attention network with structured sparsity and low-rank attention approximation.

Reference methods included a transformer-heavy reconstruction model and a hybrid convolution–attention baseline without sparsity or approximation. All models were trained with the same data splits, loss functions, and optimization settings. This setup ensures that performance differences reflect architectural design rather than training conditions.

2.3 Measurement Procedures and Quality Control

Reconstruction quality was evaluated using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). Computational efficiency was measured by peak GPU memory usage and inference throughput. All experiments were conducted on the same hardware platform with identical batch sizes. Input images were normalized to a fixed intensity range before inference. Each efficiency measurement was repeated multiple times, and average values were reported to reduce variability. Runs affected by system instability were excluded.

2.4 Data Processing and Model Formulation

Input slices were normalized and processed using lightweight convolution–attention blocks. Let $Q,K,V \in R^{N\times d}$ denote query, key, and value matrices. Standard attention has computational cost proportional to N^2, while the proposed low-rank formulation reduces this cost to

$$O(Nrd), \quad r \ll N.$$

Structured sparsity further limits computation by applying a fixed binary mask M to attention weights:

$$Y=(QK^{T} \odot M)V.$$

Reconstruction accuracy was quantified using PSNR:

$$PSNR=10\log_{10}\left(\frac{I_{max}^2}{\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{x}_i)^2}\right),$$

where x_i and $\hat{x}_i$ denote reference and reconstructed pixel values.

2.5 Evaluation Metrics and Statistical Analysis

Results are reported as mean values with standard deviations across the test set. Memory usage and inference speed were summarized relative to the transformer-heavy baseline. Statistical comparisons between models were performed using paired tests with a significance threshold of $p<0.05$. This evaluation protocol allows balanced assessment of reconstruction accuracy and computational efficiency.

3.Results and Discussion

3.1 Reconstruction Accuracy Under Efficiency Constraints

Across the two CT datasets comprising 40,000 slices, the proposed lightweight convolution–attention model preserved reconstruction accuracy close to that of the transformer-heavy baseline. The PSNR difference remained within 0.2–0.4 dB, indicating that reducing attention complexity did not lead to noticeable degradation in

overall image quality. Minor differences were mainly observed in low-contrast regions, where long-range context contributes less to reconstruction than local features [19,20]. These results show that low-rank attention and structured sparsity can reduce model complexity while maintaining the main benefits of attention-based reconstruction (Fig.1).

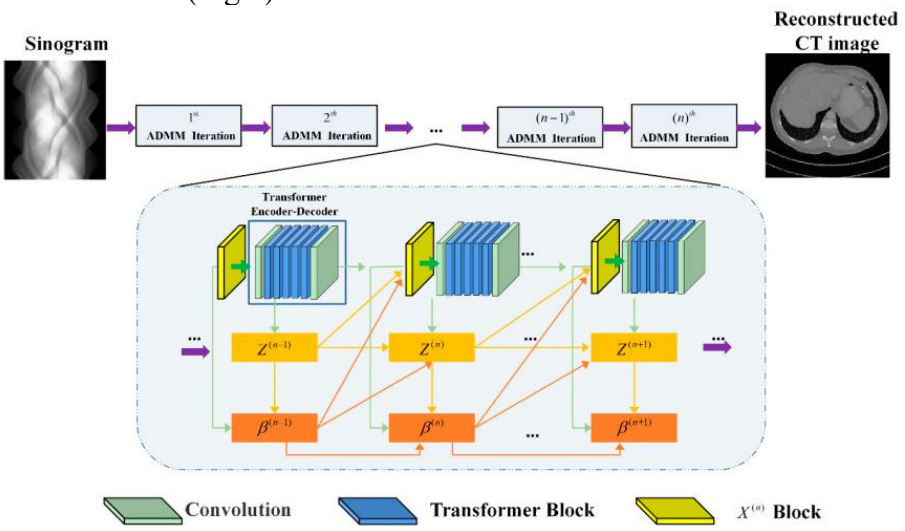

Figure 1. *Comparison of reconstruction accuracy between the lightweight convolution–attention model and transformer-heavy baselines under identical test conditions.*

3.2 Memory Usage and Inference Throughput

Substantial gains were observed in memory usage and inference speed. The proposed model reduced peak GPU memory consumption by 28%–40% and increased inference throughput by 1.6×–2.1× compared with transformer-dominant models. Memory reduction was mainly due to structured sparsity, which decreased intermediate activation storage, while throughput improvement was driven by low-rank attention that lowered attention computation cost [21]. The efficiency gains became more pronounced as image resolution increased, which is relevant for clinical deployment where large volumes are processed continuously (Fig.2).

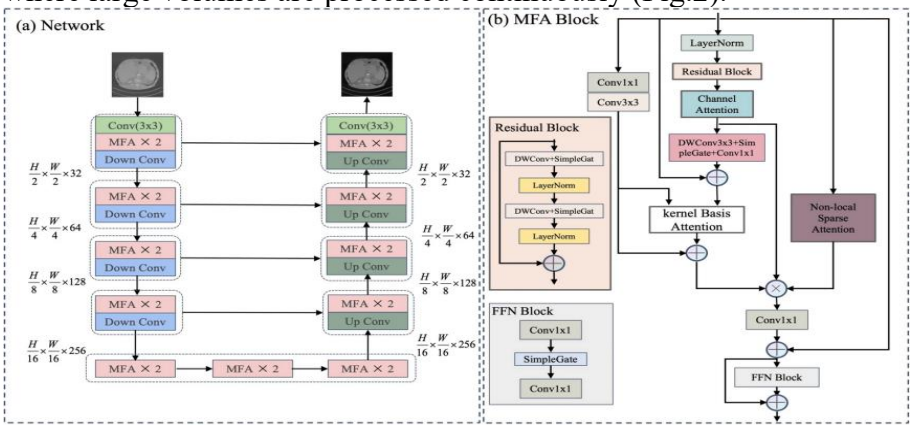

Figure 2. *GPU memory usage and inference throughput comparison for efficient and baseline CT reconstruction models.*

3.3 Comparison with Existing Efficient Reconstruction Approaches

Previous work on efficient CT reconstruction often limits attention to local windows or applies attention selectively at specific stages. While these strategies reduce computation, they also restrict the effective context available to the model. In contrast, the proposed approach retains global information flow through low-rank attention while reducing computational burden. The results suggest that reducing attention rank and enforcing structured sparsity is a more flexible strategy than spatially restricting attention, as it preserves long-range dependency modeling that is useful for suppressing structured artifacts [22,23].

3.4 Practical Implications and Limitations

The proposed method is suitable for deployment scenarios where memory and throughput directly affect clinical workflow, such as shared GPU environments and high-throughput scanners. However, aggressive rank reduction or excessive sparsity can reduce sensitivity to subtle image details, especially near high-contrast boundaries. In addition, efficiency gains depend on hardware characteristics, and sparsity patterns optimized for one platform may not transfer directly to another. Future work will focus on hardware-aware sparsity design, adaptive rank selection based on input size, and broader validation across different scanners and reconstruction settings.

4.Conclusion

This study proposes an efficient CT reconstruction method based on lightweight convolution–attention blocks with structured sparsity and low-rank attention. Results from two datasets show that the method reduces GPU memory usage and increases inference speed while keeping reconstruction accuracy close to transformer-based models. By reducing the main computational cost of attention, the method retains long-range information without dense attention operations. The findings indicate that careful architectural design can balance reconstruction quality and computational efficiency. The proposed approach is suitable for high-throughput reconstruction and shared hardware environments in clinical settings. However, reconstruction performance depends on the choice of attention rank and sparsity level, and these parameters may need adjustment for different protocols or hardware platforms. Future work will focus on adaptive parameter selection and further validation across scanners and acquisition conditions [24].

References

Kaur, N., & Brar, G. S. (2025). Advanced Methods and Approaches in Image Reconstruction. Biomedical Imaging Technology: Signal Processing Strategies and Innovations, 45-73.

Amirian, M., Barco, D., Herzig, I., & Schilling, F. P. (2024). Artifact reduction in 3D and 4D cone-beam computed tomography images with deep learning: a review. Ieee Access, 12, 10281-10295.

Wu, C., Zhu, J., & Yao, Y. (2025). Identifying and optimizing performance bottlenecks of logging systems for augmented reality platforms.

Ahmad, M., Khan, A. M., Mazzara, M., Distefano, S., Roy, S. K., & Wu, X. (2022). Hybrid dense network with attention mechanism for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 3948-3957.

Zheng, Z., Wu, S., & Ding, W. (2025). CTLformer: A Hybrid Denoising Model Combining Convolutional Layers and Self-Attention for Enhanced CT Image Reconstruction. arXiv preprint arXiv:2505.12203.

Rodríguez-Lira, D. C., Córdova-Esparza, D. M., Terven, J., Romero-González, J. A., Alvarez-Alvarado, J. M., González-Barbosa, J. J., & Ramírez-Pedraza, A. (2025). Recent developments in image-based 3D reconstruction using deep learning: Methodologies and applications. Electronics, 14(15), 3032.

Al Habsi, O., Sali, S. M., Meribout, A., Meribout, M., Almazrouei, S., & Seghier, M. (2025). Hardware acceleration in portable MRIs: State of the art and future prospects. IEEE Access.

Gonçalves, T., Rio-Torto, I., Teixeira, L. F., & Cardoso, J. S. (2022). A survey on attention mechanisms for medical applications: are we moving toward better algorithms?. IEEE Access, 10, 98909-98935.

Wang, Y., Chen, J., Arias, R., Wang, Y., & Yin, X. (2026). Development and Validation of a Patient-Friendly Digital Assessment Platform for Precision Screening of Oral Anti-Obesity Medications (AOMs).

Arora, S., Eyuboglu, S., Zhang, M., Timalsina, A., Alberti, S., Zinsley, D., ... & Ré, C. (2024). Simple linear attention language models balance the recall-throughput tradeoff. arXiv preprint arXiv:2402.18668.

Hanson, E., Li, S., Li, H. H., & Chen, Y. (2022, June). Cascading structured pruning: enabling high data reuse for sparse dnn accelerators. In Proceedings of the 49th Annual International Symposium on Computer Architecture (pp. 522-535).

Pham, N. S., Shin, S., Xu, L., Shi, W., & Suh, T. (2025). Cross-Filter Structured Pruning for Efficient Sparse CNN Acceleration. IEEE Access.

Wang, Y., Wang, Y., Yin, X., Arias, R., & Chen, J. (2026). Research on Dynamic Assessment of Glucose-Lipid Metabolism and

Personalized Drug Response Prediction Based on Wearable Multimodal Sensing.

Kwon, W., Kim, S., Mahoney, M. W., Hassoun, J., Keutzer, K., & Gholami, A. (2022). A fast post-training pruning framework for transformers. Advances in Neural Information Processing Systems, 35, 24101-24116.

Ye, M., Liu, W., Yan, L., Cheng, S., Li, X., & Qiao, S. (2021). 3D-printed Ti6Al4V scaffolds combined with pulse electromagnetic fields enhance osseointegration in osteoporosis. Molecular Medicine Reports, 23(6), 410.

Karthikeyan, V., Praveen, S., & Nandan, S. S. (2025). Lightweight deep hybrid CNN with attention mechanism for enhanced underwater image restoration. The Visual Computer, 41(8), 6251-6269.

Ahmad, M., Khan, A. M., Mazzara, M., Distefano, S., Roy, S. K., & Wu, X. (2022). Hybrid dense network with attention mechanism for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 3948-3957.

Gui, H., Zong, W., Fu, Y., & Wang, Z. (2025). Residual Unbalance Moment Suppression and Vibration Performance Improvement of Rotating Structures Based on Medical Devices.

Senthil Anandhi, A., & Jaiganesh, M. (2025). An enhanced image restoration using deep learning and transformer based contextual optimization algorithm. Scientific Reports, 15(1), 10324.

Liu, W., Zhang, W., & Ye, M. (2024). Association between carbohydrate-to-fiber ratio and the risk of periodontitis. Journal of Dental Sciences, 19(1), 246-253.

Saxena, U., Saha, G., Choudhary, S., & Roy, K. (2024). Eigen attention: Attention in low-rank space for kv cache compression. arXiv preprint arXiv:2408.05646.

Gui, H., Fu, Y., Wang, Z., & Zong, W. (2025). Research on Dynamic Balance Control of Ct Gantry Based on Multi-Body Dynamics Algorithm.

Wödlinger, M. G. (2025). Applications of Neural Attention for Modelling Long-Range Dependencies (Doctoral dissertation, Technische Universität Wien).