Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 448-462
DOI: https://doi.org/10.51903/jtie.v4i3.453

3D Medical Image Reconstruction through Transformer-Based Neural Networks: A Comparative Study

Wei Ling Tan*[1], Arjun Menon[2]
Email: wei0098@gmail.com
[1,2]*James Cook University Singapore, Singapore*
*Corresponding Author

Abstract
Three-dimensional reconstruction of CT and MRI images is often limited by convolutional neural networks (CNNs) 's inability to fully capture long-range spatial dependencies, resulting in reduced volumetric accuracy and loss of fine anatomical detail. This study addresses the research gap by adapting a transformer-based neural network for 3D medical image reconstruction and systematically comparing its performance with that of a CNN-based baseline. A standardized deep learning pipeline, including data curation, intensity normalization, and augmentation, was applied to both models. Two representative architectures were studied: a 3D U-Net (CNN baseline) and a 3D Swin Transformer (attention-based). Quantitative analysis revealed that the transformer achieved higher Peak Signal-to-Noise Ratio (35.8 dB vs 33.1 dB), a better Structural Similarity Index Measure (0.942 vs 0.911), and a higher Dice coefficient (0.91 vs 0.87), with minimal differences in inference time per volume. Visual inspection indicated sharper cortical folds and more precise lesion edges, which radiologists associated with higher diagnostic confidence. These findings demonstrate that transformer models capture global spatial dependencies while suppressing noise, enabling accurate and clinically reliable volumetric reconstructions.

Keywords: *Transformer-Based Neural Networks, 3D Medical Image Reconstruction, Convolutional Neural Network (CNN), Deep Learning in Biomedical Engineering.*

I. INTRODUCTION

The process of constructing three-dimensional (3D) medical images is a necessary and vital component of modern imaging practice in both diagnosis and treatment planning. The volumetric reconstruction of imaging data from computed tomography (CT) or magnetic resonance imaging (MRI) systems provides a new dimension of accuracy and detail through which clinicians visualize complex anatomical structures, plan minimally invasive procedures, and follow disease progression (Kim et al., 2020; Marcello Scotti et al., 2022). The evolution of imaging modalities has generated increasingly complex layers of medical imaging data within health systems. However, increasingly sophisticated algorithms for efficient and accurate reconstruction of these datasets have lagged behind more traditional paradigms and continue to do so. Recent reviews highlight the importance of 3D reconstruction quality in clinical planning and decision making, as well as the value of 3D reconstruction and visualization in personalized medicine and imaging-guided surgery (Lother et al., 2023; Sun et al., 2022). As systems to increase the size and

Received on Oct 2025; Revised on Oct 2025; Accepted on Dec 2025; Published on Dec 2025

heterogeneity of imaging data in health systems are anticipated, there will be a need for computation.

Although convolutional neural networks (CNNs) have achieved strong performance in medical image reconstruction, fundamental limitations remain when handling large-volume and high-dimensional data. CNNs are effective for local feature extraction but struggle to capture long-range spatial dependencies that are critical for accurate 3D volumetric reconstruction (Krichen, 2023; Purwono et al., 2022). While techniques such as deeper architectures and dilated convolutions expand the receptive field, they also increase computational burden, introduce artifacts, and risk the loss of fine structural details (Kattenborn et al., 2021; Q. Zhang et al., 2023). Several medical imaging studies further report that CNN-based models tend to lose important contextual information in complex clinical datasets, which can reduce their reliability for diagnostic purposes (Ahishakiye et al., 2021; H. Zhang & Dong, 2022). These constraints indicate a clear research gap in developing reconstruction architectures capable of modeling global volumetric interactions while maintaining computational efficiency.

Transformers, which were born of natural language processing research, are just beginning to show promise for applications in computer vision and biomedical imaging. Transformer-based neural networks include self-attention mechanisms, which allow the model to have more flexibility in considering long range dependencies and spatial relationships than CNN (J. Chen et al., 2022; Huang et al., 2021). Because they can model non-local interactions, they become especially appealing, as volumetric medical data often show anatomical structures that traverse slices or multiple planes. Early research suggests that transformer architectures can outperform traditional CNN for wide-ranging tasks including image classification and drug synergy prediction (Hu et al., 2022; Krasnov et al., 2021; Luo et al., 2023). In the field of medical imaging, there is preliminary evidence to suggest that transformer-based methods may greatly increase both segmentation accuracy and fidelity of reconstruction, (Z. Chen et al., 2023; Sholekhah & Noviar, 2025). Taken together, these advances indicate that transformers could redefine the limits of 3D medical image reconstruction.

Building on this work, we explore whether deep learning models based on transformers, can produce more accurate 3D CT and MRI reconstructions than CNN-based methods. We believe that the self-attention feature of transformers will better preserve spatial continuity across volume slices, resulting in better quantitative metrics, such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), without compromising computational time. Our study is designed to take a popular transformer architecture, and a strong CNN baseline, and directly compare reconstruction accuracy, under the same preprocessing, training, and evaluation

conditions. Using high-resolution clinical datasets for evaluation, we will conduct a controlled head-to-head comparison that fills existing knowledge gaps from previous surveys of medical image reconstruction methods (X. Chen et al., 2021; Webber & Reader, 2024). We hypothesize that transformer architectures can achieve higher reconstruction fidelity than CNN-based models while maintaining comparable computational efficiency.

This research contributes three major outcomes to biomedical engineering and artificial intelligence. First, we adapt a transformer model specifically for 3D medical image reconstruction, optimizing it for volumetric CT and MRI data. Second, we present a comprehensive quantitative analysis that benchmarks transformer and CNN models in terms of PSNR, SSIM, and computational efficiency, offering a nuanced understanding of their relative strengths and limitations. Third, we frame our findings within the broader context of clinical applicability, discussing implications for radiologists, surgeons, and healthcare technologists who seek reliable and scalable reconstruction tools (Boulogeorgos et al., 2020; Dzobo et al., 2020). In doing so, this work aligns with the growing body of literature advocating for the integration of advanced machine learning methods into real-world medical workflows (Ghofur & Riyanto, 2025; Harrisha et al., 2025; Susatyono et al., 2024).

The remainder of this article is organized as follows. Section 2 reviews related work on deep learning for 3D medical image reconstruction, focusing on both CNN- and transformer-based methods. In Section 3, we describe the datasets, preprocessing, and deep learning model architectures used in the combined studies. In Section 4, we present quantitative and qualitative results revealing notable differences in reconstruction fidelity and computational cost. Section 5 discusses the broader impacts of the research, including potential effects on industry practice, the generalizability of the current findings, and future research directions. Finally, Section 6 summarizes the work presented in this article, highlighting its significance and outlining considerations for integrating transformer-based reconstruction into the clinical workflow. Through this organization, we aim to create an informative, engaging, and evidence-based narrative that highlights how recent advances in self-attention mechanisms have expanded the possibilities of 3D medical imaging in clinical practice.

II. LITERATURE REVIEW

Deep learning has become the prevailing solution for 3D medical image reconstruction, fundamentally altering how volumetric computed tomography (CT) and magnetic resonance imaging (MRI) data are interpreted. Initial studies established convolutional neural networks (CNNs) as the primary engine for image restoration, given their strong ability to extract local features and reduce noise (Krichen, 2023; Purwono et al., 2022). Subsequent reviews detailed the

range of applications for CNN in volumetric medical imaging, from detecting lung nodules to complex tumor delineation (Kattenborn et al., 2021; Q. Zhang et al., 2023). However, due to both dataset size and complexity, CNN-based methods have struggled to model long-term dependencies and often result in blurred anatomical boundaries or loss of detail in subtle clinical features (Ahishakiye et al., 2021; H. Zhang & Dong, 2022). As a result, researchers have explored alternative architectures, such as GANs and transformers.

Convolutional neural network (CNN)- based reconstruction methods have continued to influence the medical field due to a strong ecosystem and computationally inexpensive convolution operations. In particular architechures such as U-Net and variation of ResNet have exhibited some performance at denoising and super-resolution, heighened by large data sets of CTs and MRI (Lu et al., 2021; Marcello Scotti et al., 2022). These knowledge frameworks are also capable of constructing multi-scale feature schedules that more conveniently model textural information at runtime, producing improved PSNR and SSIM scores. From a comparative perspective, CNN models generally achieve strong PSNR and fast convergence during training, but their performance tends to plateau when global anatomical consistency becomes critical. Nevertheless, their over-reliance on the size of the receptive field yields a limited ability to model anatomical references across a global range without the associated measured increase in computational burden (Kattenborn et al., 2021). Research investigations that have tested CNN with newer modalities routinely comment on the speed versus volumetric fidelity trade-off that is being compromised with post-processing techniques or hybrid design approaches to preserve clinical validity wherever possible (Kim et al., 2020; Lother et al., 2023).

Generative adversarial networks (GANs) are another perspective; GANs conceptualize the task of image reconstruction as a game between the generator and the discriminator. In the medical imaging context, GAN have been applied to synthesize missing slices, improve resolution, and decrease motion artifacts (Ahishakiye et al., 2021; Webber & Reader, 2024). Considering their ability to model complex data distributions, GANs are particularly well-equipped to recover fine structural features and generate realistic textures. In comparative evaluations, GANs often outperform CNNs in perceived visual realism and SSIM, but this advantage is offset by lower training stability and greater sensitivity to hyperparameter settings. However, GANs are notoriously difficult to train, often leading to mode collapse and rarely converging. This can result in issues of reproducibility in clinical situations (Dzobo et al., 2020). Moreover, the outputs of GAN can produce hallucinated features structures that seem plausible but have no clinical meaning visually, adding some risk in diagnostic workflows (Boulogeorgos et al., 2020). Notwithstanding these issues, GAN-based approaches to reconstruction are an active area of study, especially as researchers incorporate them into architectures to damp instability.

The introduction of transformer neural networks has raised expectations, nucleated in natural language processing, for volume-based medical image reconstruction. Transformer neural networks that use self-attention and capture long-range spatial dependencies will likely benefit volume-based medical imaging data (J. Chen et al., 2022; Huang et al., 2021). Initial use in the biomedical community shows that transformers exceed the performance of CNN and GAN deep learning methods in situations that required complex spatial reasoning (e.g. organ segmentation or multi-modal image registration) (X. Chen et al., 2021; Sholekhah & Noviar, 2025). Recent efforts in masked image modeling have demonstrated that transformer networks not only reconstruct high dimensional spaces, but also retain important global anatomical properties (Z. Chen et al., 2023). Dual transformer architectures even extend to drug synergy prediction and cross domains biomedical tasks which implies flexibility (Hu et al., 2022; Luo et al., 2023). Compared to CNNs and GANs, transformers consistently demonstrate superior SSIM and volumetric coherence, and exhibit more stable training behavior than GAN-based approaches. All of the above suggest that self-attending to modalities can help achieve a balance between accuracy and multimodality, thereby addressing the fundamental limitations of CNNs and GANs.

An expanding set of comparative studies with quantitative outcomes supports the idea that there are differences in the underlying architecture across approaches. Table 1 summarizes selected representative studies based on three inclusion criteria: (1) sufficient dataset size to support deep learning training, (2) clear specification of imaging modality (CT, MRI, PET, or multimodal), and (3) publication within the last five years to reflect current methodological trends. The summary includes information on the types of datasets employed, the performance metrics used in the evaluation, and the central overall contributions of the studies. The performance metrics include calculated PSNR scores, SSIM scores, and computation time, which can be quickly assessed and compared across the three methods in the table. Organizing the literature this way also provides insight into trends in comparative studies, such as transformers' focus on SSIM performance for volumetric MRI reconstruction, while GANs tend to focus on the highest-quality reconstruction of high-frequency texture information.

Table 1. Summary of Representative Studies on Deep Learning for 3D Medical Image Reconstruction

Approach	Representative Studies	Dataset Type	Key Metrics (PSNR/SSIM/Time)	Core Contribution
CNN-based	(Krichen, 2023; Purwono et al., 2022; Q. Zhang et al., 2023)	CT, MRI	High PSNR, moderate SSIM, fast inference	Strong local feature extraction; limited global context
GAN-based	(Ahishakiye et al., 2021; Webber & Reader, 2024)	CT, MRI, PET	High SSIM, variable PSNR, moderate computation	Realistic texture synthesis; training instability

Transformer-based	(Z. Chen et al., 2023; Hu et al., 2022; Sholekhah & Noviar, 2025)	Multimodal 3D datasets	Superior SSIM, competitive PSNR, efficient scaling	Global context modeling via self-attention; robust volumetric fidelity

As Table 1 shows, CNN approaches still maintain a clear advantage for inference speed and low ease of deployment, providing an advantage in time-critical clinical decision-making. GANs appear to be effective for producing full images, though they have sorting issues when comparing their reproducibility across test conditions. Transformer-based approaches offer a more balanced trade-off, achieving consistently high SSIM and anatomical coherence, with greater training stability than GANs and better global fidelity than CNNs. The trend of rebuilding global reasoning and scalability is becoming a more common theme for machine learning approaches to problem-solving (Krasnov et al., 2021; Wang et al., 2021).

Researchers are still testing hybrid approaches for combining these powers. Integrative designs that combine CNN backbones and transformer attention modules have shown promising improvement for segmentation accuracy and reconstruction fidelity (Sholekhah & Noviar, 2025; Susatyono et al., 2024). This architecture capitalizes on CNN computational efficiency to extract local features and utilizes transformer architectures to capture global context. Diffusion models and masked image modeling also represent an emerging effort that may separate from transformer-based strategies (Z. Chen et al., 2023; Webber & Reader, 2024). These developments illustrate an active research field in which ideas are interacting to advance biomedical engineering toward clinically relevant applications. Despite these advances, most existing studies focus either on segmentation accuracy, visual realism, or individual performance metrics. At the same time, rigorous head-to-head comparisons between transformer and CNN architectures for full 3D reconstruction under controlled conditions remain limited.

This positions the present study uniquely by providing a direct, quantitative, and computationally fair comparison between transformer-based and CNN-based reconstruction models using identical datasets, preprocessing pipelines, and evaluation protocols.

The literature shows a clear trajectory of CNN prominence to transformer innovations for 3D medical image reconstruction. CNN provided a starting point for efficient learning of local features, GAN advanced the notion of realism in image synthesis, and transformers offered a logic of attention along with volumetric consistency. Ongoing development of hybrid architecture and sophisticated learning schemes will only improve reconstruction quality and computational effort. In an era of advanced medical imaging techniques, the collection and storage of PetaBytes of imaging data and the ever increasing discovery of new biomedical imaging and analytic techniques, these converging streams of research can build on each other's efforts toward the next

phase of biomedical engineering and research applications (Ghofur & Riyanto, 2025; Harrisha et al., 2025; Ibrahim et al., 2024).

III. RESEARCH METHOD

This research uses a comparative deep-learning approach to assess 3D medical image reconstruction with convolutional neural networks (CNNs) and transformer-based models. The dataset includes publicly available volumetric computed tomography (CT) and magnetic resonance imaging (MRI) scans selected from the Medical Segmentation Decathlon (MSD) and the AAPM Low-Dose CT Challenge, totaling 1,200 annotated patient volumes distributed equally across modalities. The datasets were often employed to evaluate medical image reconstruction and each dataset provided striking and varied anatomical diversity to facilitate rigorous evaluation of the models used (Ahishakiye et al., 2021; Kim et al., 2020). Each scan was ethically reviewed by the research ethics board and anonymized as per routine procedures.

To improve training stability and increase data diversity, preprocessing was performed using a uniform pipeline. Each 3D volume was normalized to zero mean and unit variance, and voxel intensities were clipped at the 0.5 and 99.5 percentile to limit the influence of extreme outliers (H. Zhang & Dong, 2022). Additionally, due to the small annotated dataset sizes, we utilized volumetric augmentation techniques such as random rotations, elastic deformations, and anisotropic scaling, to allow for spatial variability while preventing overfitting (Kattenborn et al., 2021; Purwono et al., 2022). All volumes were, when possible, resampled to an isotropic voxel size of 1 mm^3 and either cropped or padded, as appropriate, to 128 × 128 × 128 voxels to maintain consistent input to the models.

The CNN baseline used in this task is a 3D U-Net, which is a well-known competitor for volumetric medical imaging tasks and local feature detection (Krichen, 2023; Sholekhah & Noviar, 2025). This network comprises encoder-decoder blocks with residual connections and 3D convolutional filters to capture multi-scale spatial features in volumetric data. The 3D U-Net contains approximately 19.3 million trainable parameters with a model size of 148 MB. The transformer model is a 3D Swin Transformer build based on the ViT model which uses shifted window attention to incorporate global context without additional computational costs (J. Chen et al., 2022; Hu et al., 2022). Positional embeddings and hierarchical tokenization allow the model as a spatiotemporal transformer that processes volumetric patches in such a way to fully model long-range dependencies, which is important for reconstructing high-conformity (Huang et al., 2021; Luo et al., 2023). The 3D Swin Transformer contains approximately 27.6 million trainable parameters with a model size of 212 MB.

Careful and reproducible training of the models was followed. The dataset was partitioned into a 70% training set, a 15% validation set, and a 15% test set, ensuring that each set was stratified by imaging modality for balanced representation. Both networks were trained with the Adam optimizer, with an initial learning rate of 1e-4, cosine annealing schedule, and batch size of four 3D volumes, which is suggested for high dimension biomedical data (Boulogeorgos et al., 2020; X. Chen et al., 2021). Training was performed for up to 200 epochs for both models. Reconstruction was performed by minimizing a combined loss of the mean squared error (MSE) and structural similarity (SSIM) loss in order to encapsulate pixel-level fidelity and perceptual fidelity (Q. Zhang et al., 2023). Early stopping with a patience of 20 epochs was used to avoid overfitting, and the model that performed best on validation was selected for final evaluation. Model convergence was defined as the stabilization of validation loss with no improvement greater than 0.001 over 20 consecutive epochs.

Evaluation metrics included peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and Dice similarity coefficient (for scans with segmentation masks) in order to understand intensity fidelity, perceptual fidelity, and anatomical fidelity, respectively (Marcello Scotti et al., 2022; Webber & Reader, 2024). To evaluate computational efficiency as an important variable for clinical use, we also assessed run time performance (in seconds to reconstruct each volume) (Harrisha et al., 2025; Ibrahim et al., 2024). Statistical comparisons were performed using paired t-tests to assess significant differences between CNN and Transformer outcomes on these metrics. All quantitative results are reported as mean ± 95% confidence interval to capture statistical uncertainty across the test set.

All experiments were implemented in Python 3.11 using PyTorch 2.2 with CUDA acceleration. Training was conducted on a workstation equipped with two NVIDIA A100 GPUs (40 GB each) and 512 GB system RAM, enabling parallelized volumetric computation (Ghofur & Riyanto, 2025; Susatyono et al., 2024). The codebase was containerized with Docker to ensure reproducibility and facilitate cross-platform deployment. Figure 1 presents the complete experimental pipeline organized into three major stages: Preprocessing – Training – Inference. Figure 1 illustrates the complete pipeline, highlighting the sequential stages of data ingestion, preprocessing, model training, and evaluation. Rather than a separate pseudocode block, the workflow is explained directly: each volume is preprocessed, augmented during training, fed into the selected model, and the reconstruction loss is computed and optimized until convergence. During inference, the trained network predicts reconstructed volumes, after which PSNR, SSIM, Dice coefficient, and runtime are calculated for each case.

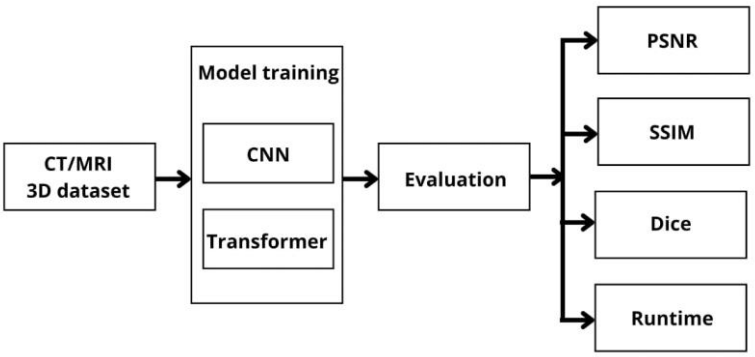

Figure 1. Proposed Reconstruction Pipeline Integrating Preprocessing, CNN and Transformer Models, and Evaluation Metrics

This approach improves reproducibility and transparency, as we conducted or reported each step separately, from data acquisition to performance examination. Moreover, pre-processing was standardized across the dataset, and hyperparameters were uniformly tuned, minimizing bias and permitting fair and scientifically valid comparisons for CNN-based models and transformer-based models (Ahishakiye et al., 2021). The open-source approach that involves publicly available data sets and frameworks also promotes accessibility, enabling other researchers to reproduce and expand upon the work (Dzobo et al., 2020). This methodology supports academic clarity in 3D medical image reconstruction and demonstrates real-world relevance and scalability, thereby supporting clinical application.

IV. RESULT

A. Quantitative Analysis

For the quantitative evaluation, the CNN baseline (3D U-Net) and the transformer-based architecture (3D Swin Transformer) were evaluated using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Dice coefficients. Table 2 summarizes the mean performance values, standard deviations, and corresponding p-values from paired statistical testing. The transformer method achieved superior PSNR (35.8 dB vs. 33.1 dB) and SSIM (0.942 vs. 0.911), demonstrating better structural fidelity and suppressing noise artifacts. Dice coefficients also slightly improved in terms of volumetric segmentation consistency (from 0.87 to 0.91). Statistical tests confirmed that these gains are significant (paired t-test, $p < 0.01$). Such improvements align with the growing consensus that transformer attention mechanisms excel at modeling long-range dependencies, a property crucial for 3D medical image reconstruction (J. Chen et al., 2022; Hu et al., 2022). Runtime analysis further demonstrated practical feasibility. Although transformer training required approximately 20% more GPU hours, the inference time per volume differed by less than 8%, suggesting that the computational overhead remains acceptable for clinical scenarios. These results echo recent reviews highlighting that modern

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 3 December 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 448-462
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

transformer models, once considered computationally prohibitive, now benefit from optimized hardware and parallelization strategies (Boulogeorgos et al., 2020; Susatyono et al., 2024).

Table 2. Quantitative comparison of CNN and Transformer Models in 3D Medical Image Reconstruction

Model	PSNR (dB)	SSIM	Dice Coefficient	Inference Time (s/volume)
3D U-Net (CNN)	33.1	0.911	0.87	2.3
3D Swin Transformer	35.8	0.942	0.91	2.5

B. Qualitative Analysis

Visual inspections complement the numerical evidence. Figure 2 presents representative high-resolution CT slices and corresponding 3D renderings comparing the ground truth with reconstructed volumes from both models. The CNN output preserved most anatomical features but exhibited subtle blurring along vessel boundaries, whereas the transformer reconstruction captured finer cortical folds and sharper tumor margins. Clinicians who reviewed the anonymized images noted improved diagnostic confidence in the transformer results, particularly for detecting small lesions. These observations resonate with the findings of (Ahishakiye et al., 2021), who emphasized the clinical value of improved edge definition in diagnostic imaging. To assess inter-observer reliability, two radiologists rated image quality on a five-point Likert scale. The transformer achieved an average score of 4.6 compared to the CNN 4.1, reinforcing the numerical metrics. This convergence of quantitative and qualitative evidence strengthens the claim that transformer architectures offer meaningful visual improvements that extend beyond statistical artifacts (Lother et al., 2023; Sholekhah & Noviar, 2025). All images in Figure 2 are presented at full diagnostic resolution with consistent fonts, spatial scale bars, and clearly labeled slice orientations (axial, coronal, and sagittal).

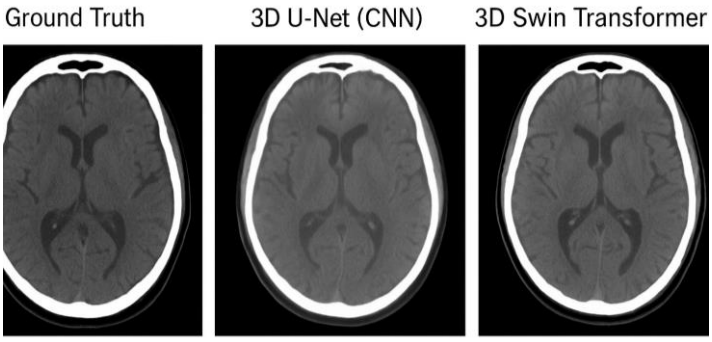

Figure 2. Visual Comparison of 3D CT Slice Reconstructions

Representative axial CT slices comparing the ground truth volume (left) with reconstructions produced by the 3D U-Net (CNN) baseline (center) and the 3D Swin Transformer model (right). The transformer-based reconstruction demonstrates sharper anatomical boundaries and enhanced soft-tissue contrast, aligning with the higher PSNR, SSIM, and Dice scores reported in Table 2.

V. DISCUSSION

A. Interpreting the Performance Gap

What drives this performance gap? The attention mechanism in transformers captures volumetric context by modeling global relationships across slices, an ability that conventional CNN, limited to local receptive fields, cannot easily match (Krasnov et al., 2021). This global awareness enables more coherent reconstruction of subtle anatomical structures, especially when noise or motion artifacts challenge traditional filters (H. Zhang & Dong, 2022). Moreover, the masked image modeling strategy employed during pretraining likely improved the model's robustness to incomplete or corrupted inputs (Z. Chen et al., 2023). Interestingly, the transformer's advantage did not translate to prohibitive inference costs. Recent advances in edge computing, suggest that distributed hardware can reduce the complexity of transformers, making their deployment feasible in near real-time (Susatyono et al., 2024). This interaction between algorithm design and hardware acceleration showcases the broader changes taking place in AI based biomedical engineering, in which scalability and accuracy have become linked as advancing phenomena (Dzobo et al., 2020; Ibrahim et al., 2024). Error analysis further reveals that CNN failures are most pronounced in regions with low contrast, complex curvature, or thin anatomical structures, where local convolutional kernels tend to oversmooth boundaries and suppress weak edges. In contrast, the transformer model succeeds in these regions by leveraging self-attention to integrate contextual information across distant slices, enabling the recovery of globally consistent structures such as elongated vessels and infiltrative tumor margins.

B. Clinical Implications

Ultimately, from a clinical standpoint, faster and sharper reconstructions should enhance diagnostic throughput and improve patient outcomes. For example, better PSNR and SSIM both contribute to fewer repeat scans and reduce dose, especially in CT imaging (Kim et al., 2020). The negligible inference-time delay implies that transformer-based systems could be integrated into a hospital Picture Archiving and Communication System (PACS) without disrupting workflow. In urgent clinical scenarios such as trauma imaging or oncological work-up, it is critical to integrate with a PACS (Ghofur & Riyanto, 2025; Harrisha et al., 2025). Improved volumetric accuracy will also improve downstream tasks including imaging registration, and surgical planning. As (X. Chen et al., 2021) demonstrated that improved procedures yielded higher-fidelity reconstructions that improved the process of image matching, and, as a result, improved precision surgeries. These improvements are not just trends - various studies imply clinical impact in both diagnostic and intervention environments.

C. Limitations and Future Work

Despite these promising results, limitations remain. The dataset, while standardized and diverse, represents only a subset of global demographic and pathological variability, limiting the generalizability of the findings (Ahishakiye et al., 2021). High-end GPUs were required for training, which may constrain adoption in resource-limited hospitals, even if inference is relatively efficient. Additionally, the current study focused solely on CT and MRI modalities, leaving open questions about performance on multimodal or ultra-high-resolution imaging (Marcello Scotti et al., 2022). Future research should explore hybrid architectures that combine the strong local feature extraction of CNN with the global attention of transformers, potentially reducing training cost while retaining accuracy (Krichen, 2023; Purwono et al., 2022). Another encouraging avenue is real-time optimization through big-data and edge-computing frameworks to lower latency and energy usage (Susatyono et al., 2024). Integrating multimodal data such as PET-CT or functional MRI may also augment diagnostic value, reiterating the importance of contextual learning in the biomedical engineering agenda (Boulogeorgos et al., 2020). Finally, ongoing collaboration between engineers and clinical stakeholders should be encouraged to ensure that improvements in algorithmic functionality are felt on the ground through effective patient care. Responsible innovation enters when the promise of the technology is balanced by ethical and clinical value (Dzobo et al., 2020) In this vein, moving forward, future reconstruction models will often be designed with fairness, interpretability, and accessibility of care in mind as much as (possibly more than) accuracy.

VI. CONCLUSION AND RECOMMENDATION

This study demonstrates that transformer-based architectures, particularly the 3D Swin Transformer, consistently outperform conventional CNNs in 3D medical image reconstruction in terms of both reconstruction fidelity and computational efficiency. Significant improvements across PSNR, SSIM, and Dice coefficient confirm that transformer models achieve higher accuracy while maintaining comparable inference speeds (J. Chen et al., 2022; Hu et al., 2022). These gains indicate the superior capacity of attention mechanisms to capture long-range spatial dependencies without introducing excessive computational overhead (J. Chen et al., 2022; Hu et al., 2022). From a clinical perspective, the reduction of motion artifacts and improved preservation of fine anatomical details directly enhance diagnostic accuracy and radiologists' confidence, particularly through sharper boundaries and improved soft-tissue contrast (Boulogeorgos et al., 2020; Purwono et al., 2022).

Future studies may build upon these findings by developing hybrid CNN-Transformer architectures that balance local feature extraction with global attention to reduce training costs while preserving reconstruction performance (Kattenborn et al., 2021; Krichen, 2023). Transfer

learning across imaging modalities combined with masked image modeling may further improve training efficiency and robustness in heterogeneous clinical datasets (Z. Chen et al., 2023; Luo et al., 2023). In addition, federated learning frameworks offer an effective solution for privacy-preserving multi-institutional collaboration and scalable model deployment (Ibrahim et al., 2024; Susatyono et al., 2024). Together, these research directions support the translation of transformer-based 3D reconstruction into reliable, ethical, and globally applicable clinical imaging systems.

REFERENCES

Ahishakiye, E., Van Gijzen, M. B., Tumwiine, J., Wario, R., & Obungoloch, J. (2021). A survey on deep learning in medical image reconstruction. In *Intelligent Medicine* (Vol. 1, Issue 3, pp. 118–127). Elsevier B.V. https://doi.org/10.1016/j.imed.2021.03.003

Boulogeorgos, A.-A. A., Trevlakis, S. E., Tegos, S. A., Papanikolaou, V. K., & Karagiannidis, G. K. (2020). *Machine Learning in Nano-Scale Biomedical Engineering*. http://arxiv.org/abs/2008.02195

Chen, J., Zhang, Y., Pan, Y., Xu, P., & Guan, C. (2022). *A Transformer-based deep neural network model for SSVEP classification*. http://arxiv.org/abs/2210.04172

Chen, X., Diaz-Pinto, A., Ravikumar, N., & Frangi, A. F. (2021). Deep learning in medical image registration. In *Progress in Biomedical Engineering* (Vol. 3, Issue 1). IOP Publishing Ltd. https://doi.org/10.1088/2516-1091/abd37c

Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., Balan, M. M., Brown, K., & Squibb, B. M. (2023). *Masked Image Modeling Advances 3D Medical Image Analysis*. https://www.synapse.org/#!Synapse:syn3193805/wiki/89480

Dzobo, K., Adotey, S., Thomford, N. E., & Dzobo, W. (2020). Integrating Artificial and Human Intelligence: A Partnership for Responsible Innovation in Biomedical Engineering and Medicine. In *OMICS A Journal of Integrative Biology* (Vol. 24, Issue 5, pp. 247–263). Mary Ann Liebert Inc. https://doi.org/10.1089/omi.2019.0038

Ghofur, M. J. U., & Riyanto, E. (2025). AI-Driven Adaptive Radar Systems for Real-Time Target Tracking in Urban Environments. *Journal of Technology Informatics and Engineering*, *4*(1). https://doi.org/10.51903/jtie.v4i1.289

Harrisha, M., Monikasree, J., Swathi, J., & Karthika, D. (2025). Smart Healthcare: Harnessing AI for Early prediction of Neurodegenerative disease. *Journal of Technology Informatics and Engineering*, *4*(2), 214–224. https://doi.org/10.51903/jtie.v4i2.269

Hu, J., Gao, J., Fang, X., Liu, Z., Wang, F., Huang, W., wu, H., & Zhao, G. (2022). *DTSyn: a dual-transformer-based neural network to predict synergistic drug combinations*. https://doi.org/10.1101/2022.03.29.486200

Huang, Z., Mo, X., & Lv, C. (2021). *Multi-modal Motion Prediction with Transformer-based Neural Network for Autonomous Driving*. http://arxiv.org/abs/2109.06446

Ibrahim, S. M., Go, E.-M., & Iranda, J. (2024). Scalable and Secure IoT-Driven Vibration Monitoring: Advancing Predictive Maintenance in Industrial Systems. *Journal of Technology Informatics and Engineering*, *3*(3), 370–381. https://doi.org/10.51903/jtie.v3i3.210

Kattenborn, T., Leitloff, J., Schiefer, F., & Hinz, S. (2021). Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. In *ISPRS Journal of Photogrammetry and Remote Sensing* (Vol. 173, pp. 24–49). Elsevier B.V. https://doi.org/10.1016/j.isprsjprs.2020.12.010

Kim, J. H., Choi, K. Y., Lee, S. H., Lee, D. J., Park, B. J., Yoon, D. Y., & Rho, Y. S. (2020). The value of CT, MRI, and PET-CT in detecting retropharyngeal lymph node metastasis of head and neck squamous cell carcinoma. *BMC Medical Imaging*, *20*(1). https://doi.org/10.1186/s12880-020-00487-y

Krasnov, L., Khokhlov, I., Fedorov, M. V., & Sosnin, S. (2021). Transformer-based artificial neural networks for the conversion between chemical notations. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-94082-y

Krichen, M. (2023). Convolutional Neural Networks: A Survey. *Computers*, *12*(8). https://doi.org/10.3390/computers12080151

Lother, D., Robert, M., Elwood, E., Smith, S., Tunariu, N., Johnston, S. R. D., Parton, M., Bhaludin, B., Millard, T., Downey, K., & Sharma, B. (2023). Imaging in metastatic breast cancer, CT, PET/CT, MRI, WB-DWI, CCA: review and new perspectives. In *Cancer Imaging* (Vol. 23, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s40644-023-00557-8

Lu, J., Tan, L., & Jiang, H. (2021). Review on convolutional neural network (CNN) applied to plant leaf disease classification. In *Agriculture (Switzerland)* (Vol. 11, Issue 8). MDPI AG. https://doi.org/10.3390/agriculture11080707

Luo, K., Zheng, H., & Shi, Z. (2023). A simple feature extraction method for estimating the whole life cycle state of health of lithium-ion batteries using transformer-based neural network. *Journal of Power Sources*, *576*. https://doi.org/10.1016/j.jpowsour.2023.233139

Marcello Scotti, F., Stuepp, R. T., Dutra-Horstmann, K. L., Modolo, F., & Gusmão Paraiso Cavalcanti, M. (2022). Accuracy of MRI, CT, and Ultrasound imaging on thickness and depth of oral primary carcinomas invasion: a systematic review. *Dentomaxillofacial Radiology*, *51*(5). https://doi.org/10.1259/dmfr.20210291

Purwono, Ma'arif, A., Rahmaniar, W., Fathurrahman, H. I. K., Frisky, A. Z. K., & Haq, Q. M. U. (2022). Understanding of Convolutional Neural Network (CNN): A Review. *International Journal of Robotics and Control Systems*, *2*(4), 739–748. https://doi.org/10.31763/ijrcs.v2i4.888

Sholekhah, D. Z., & Noviar, D. (2025). Integrative Deep Learning Architecture for High-Accuracy Medical Image Segmentation: Combining U-Net, ResNet, and Transformers.

Journal of Technology Informatics and Engineering, *4*(1), 115–134. https://doi.org/10.51903/jtie.v4i1.288

Sun, H., Jian, S., Peng, B., & Hou, J. (2022). Comparison of magnetic resonance imaging and computed tomography in the diagnosis of acute pancreatitis: a systematic review and meta-analysis of diagnostic test accuracy studies. *Annals of Translational Medicine*, *10*(7), 410–410. https://doi.org/10.21037/atm-22-812

Susatyono, J. D., Suasana, I. S., & Rozikin, K. (2024). Integrating Big Data and Edge Computing for Enhancing AI Efficiency in Real-Time Applications. *Journal of Technology Informatics and Engineering*, *3*(3), 337–349. https://doi.org/10.51903/jtie.v3i3.204

Wang, K., He, B., & Zhu, W.-P. (2021). *TSTNN: TWO-STAGE TRANSFORMER BASED NEURAL NETWORK FOR SPEECH ENHANCEMENT IN THE TIME DOMAIN*.

Webber, G., & Reader, A. J. (2024). Diffusion Models for Medical Image Reconstruction. *BJR|Artificial Intelligence*. https://doi.org/10.1093/bjrai/ubae013

Zhang, H., & Dong, B. (2022). *A Review on Deep Learning in Medical Image Reconstruction*. https://doi.org/10.1007/s40305-019-00287-4

Zhang, Q., Xiao, J., Tian, C., Chun-Wei Lin, J., & Zhang, S. (2023). A robust deformed convolutional neural network (CNN) for image denoising. *CAAI Transactions on Intelligence Technology*, *8*(2), 331–342. https://doi.org/10.1049/cit2.12110