Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 16, No. 2, 2026

Lightweight U-Net++ with Hybrid Swin-Convolutional Attention for Efficient CT Metal Artifact Reduction

Shanshan Wang, Ao Zang, Hao Zhang, Peng Lu, Jiang Ma *

North China University of Science and Technology, Tangshan, Hebei Province, 063210, China

*** Corresponding author:** Jiang Ma

Abstract: Metal artifacts from high-density implants severely degrade CT images, hindering accurate diagnosis and downstream tasks. Existing Metal Artifact Reduction (MAR) methods face a critical trade-off: CNNs struggle with global artifacts due to limited receptive fields, while Transformers incur prohibitive computational costs unsuitable for real-time clinical use. To address this, we propose U-Net++(SwinLite), an efficient, lightweight MAR architecture. By integrating a nested dense cascade skip-connection mechanism and a Hybrid Attention Module (coupling Swin and convolutional attention), our model effectively captures both local details and long-range dependencies. Using a dynamic channel adjustment strategy, U-Net++(SwinLite) achieves a Mean Absolute Error (MAE) of 9.6006 with only 3.05M parameters —reducing model size by over 95% compared to pure Transformer models like Restormer-U. Extensive experiments demonstrate our method achieves an optimal balance between structural preservation and computational efficiency.

Keywords: Computed Tomography (CT); U-Net++; Swin Transformer; Lightweight Architecture; Medical Image Segmentation.

1. Introduction

Medical imaging technologies, particularly CT and X-ray imaging, are indispensable tools in modern medical diagnostics, providing intuitive and precise anatomical visualization. However, the presence of metallic materials—such as orthopedic screws, artificial joints, and dental fillings—causes severe photon starvation and beam hardening during the imaging process. These physical phenomena manifest as starburst or streak artifacts, dark signal voids, and localized overexposure[5]. Such artifacts heavily distort adjacent anatomical structures, creating a major hurdle for clinical evaluation and severely degrading the performance of automated diagnostic pipelines, including dental panoramic X-ray image instance segmentation and 3D reconstruction systems[6].

Historically, MAR was addressed using sinogram-domain interpolation, which often resulted in secondary artifacts or the loss of subtle structural details near the implants due to over-smoothing[7]. The paradigm shift towards deep learning, particularly the U-Net architecture, significantly improved image-domain restoration by learning non-linear mappings from artifact-corrupted to clean images. Nevertheless, CNN-based MAR approaches are fundamentally constrained by their localized receptive fields, struggling to eliminate global, long-spanning streak artifacts without compromising tissue fidelity[8].

Recently, Vision Transformers have been introduced to MAR to leverage their global context modeling capabilities. Models like Restormer-U have achieved remarkable performance[9]; however, they introduce a severe computational bottleneck. In time-sensitive clinical scenarios, such as intraoperative portable CT guidance or rapid dental screenings, models are rigidly constrained by hardware (often < 8GB VRAM) and require minimal inference latency[10]. A model with over 60 million parameters is simply unfeasible for such deployments[11].

To bridge this gap, this paper introduces a highly efficient hybrid network, UNet++(SwinLite). Our core contributions are:We design the NDC-SCM to ensure that critical artifact boundary details (e.g., bone trabeculae and soft tissue textures) are not lost during the downsampling process[12].We propose a high-resolution Hybrid Attention Module that operates on non-overlapping windows, drastically reducing the quadratic complexity of global attention while maintaining structural coherence[13].We implement a dynamic channel allocation scaling factor that successfully shrinks the parameter count to 3.05M while preserving robust denoising precision.

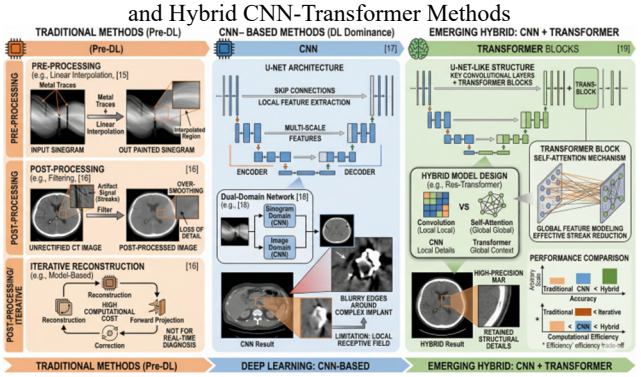
2. Related Work

Traditional Metal Artifact Reduction Before the widespread adoption of deep learning, traditional MAR methods primarily relied on mathematical modeling of the imaging physics or optimizing pixel intensity distributions[14]. Pre-processing methods typically involve identifying metal regions in the projection data and applying linear interpolation or weighted averaging to fill the missing data gaps[15]. However, these methods frequently discard structural details at the metal-tissue interface. Post-processing techniques attempt to separate artifact signals from anatomical structures in the image domain using filters, which can lead to over-smoothing. Iterative reconstruction methods refine images by cycling between image reconstruction and projection correction, but their high computational cost makes them unsuited for real-time diagnosis[16].

Deep Learning-based MAR CNNs have dominated medical image restoration due to their excellent local feature extraction capabilities[17]. U-Net and its variants utilize encoder-decoder structures with skip connections to preserve multi-scale features. While dual-domain networks (operating on both sinogram and image domains) further suppress artifacts, they remain limited by the inherent local receptive field of convolutions, often resulting in blurred edges around complex implants. Alternatively, Transformer-based architectures introduce self-attention mechanisms for global

feature modeling[18]. While effective in reducing widespread streak artifacts, pure Transformer networks suffer from weak local detail capture and prohibitive computational complexity, making the pursuit of a high-precision, high-efficiency hybrid model a pressing need [19].

Table 1. CT Metal Artifact Reduction: Traditional, CNN-Based, and Hybrid CNN-Transformer Methods

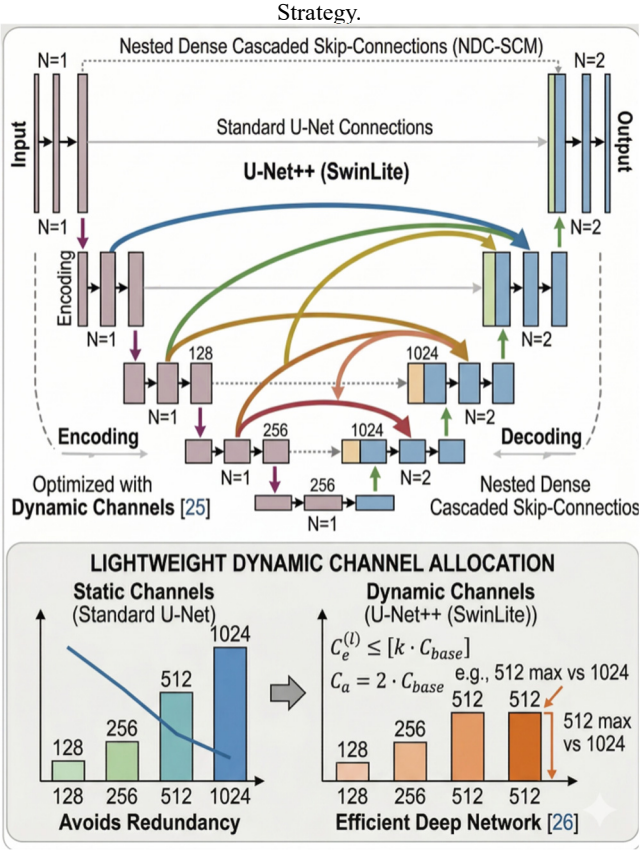


3. Methodology

3.1. Overall Architecture

The proposed U-Net++(SwinLite) is an end-to-end denoising network tailored for MAR. It builds upon the U-Net++ framework by replacing standard connections with dense cascading pathways and integrating a specialized hybrid attention mechanism. The architecture is systematically optimized for limited-resource environments through dynamic channel configuration.

Table 2. Overall Architecture and Dynamic Channel Allocation Strategy.



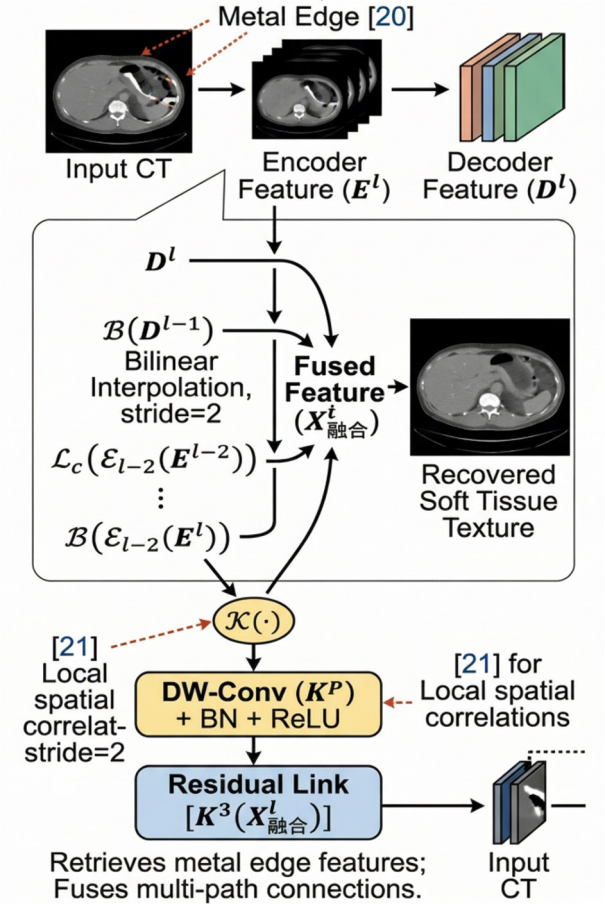
3.2. Nested Dense Cascade Skip-Connections (NDC-SCM)

Traditional U-Net++ relies on direct concatenation between encoder and decoder layers of the same level, which is insufficient for handling the complex frequency distribution of metal artifacts[20]. To recover high-frequency soft-tissue textures obscured by metallic edges, our NDC-SCM establishes multi-path connections between the encoder layer i and the decoder layer j. The feature fusion is mathematically formulated as:

$$F_{i,j} = \mathrm{Conv}_{1\times1}\big(\mathrm{Cat}(F_{i,j-1}, \mathrm{UpSample}(F_{i+1,j}))\big) \oplus \mathrm{DW}-\mathrm{Conv}(F_{i,j-1}) \quad (1)$$

Where $F_{i,j}$ denotes the fused feature map, and Cat(·) represents channel-wise concatenation. UpSample utilizes bilinear interpolation (stride = 2). The 1×1 convolution reduces dimensional redundancy, while the depth-wise convolution ($\mathrm{DW}-\mathrm{Conv}$) with a 3×3 kernel enhances local spatial correlations[21]. The residual connection $\oplus$ effectively mitigates gradient vanishing in deep networks.

Table 3. Nested Dense Cascaded Skip-Connection Module (NDC-SCM).



3.3. Hybrid Convolutional-Swin Attention Module

To effectively suppress irregular artifacts, we embed a Hybrid Attention Module in the high-resolution layers (1/2 to 1/4 image scale). The module partitions the feature map into non-overlapping 8×8 windows to restrict self-attention computation, thereby lowering global complexity[22]. For a given window, the Multi-Head Self-Attention (MSA) is computed as:

$$\mathrm{MSA}(Q,K,V) = \mathrm{Cat}_{k=1}^{K}\left(\mathrm{SoftMax}\left(\frac{Q_k K_k^T}{\sqrt{d_k}} + M_k\right)V_k\right)W^O \quad (2)$$

Where $K = 4$ represents the number of attention heads, $d_k = d_{model}/K = 16$ is the dimension per head, and M_k is the

window mask matrix preventing out-of-window pixel interactions[23]. To restore local sensitivity and achieve cross-window global association, this transformer branch is fused with a parallel convolutional attention branch via a shifted-window mechanism and dynamic gating:

$$F_{hybrid} = Conv_{3\times3}(F_{attn}) \odot \sigma(GlobalAvgPool(F_{attn})) \quad (3)$$

$$F_{attn} = LayerNorm(MSA(F_{in}) + F_{in}) \quad (4)$$

Where $\sigma(\cdot)$ is the Sigmoid activation function that adaptively activates artifact-relevant channels, ensuring local gradient sensitivity and global striation modeling[24].

3.4. Lightweight Dynamic Channel Allocation Strategy

The efficiency of U-Net++(SwinLite) is heavily attributed to its base-parameter (C_{base}) dynamic configuration. Instead of static channel doubling (e.g., 128, 256, 512, 1024) which causes severe computational redundancy[25], we restrict the encoder stage i channels according to:

$$C_{encoder}(i) = C_{base} \times 2^i \quad (5)$$

Setting the baseline parameter $C_{base} = 64$ limits the highest semantic feature layer (Stage 4) to 512 channels. Consequently, the hybrid attention modules are dynamically fixed to double the base value ($C_{attn} = C_{base} \times 2 = 128$). This design prevents the explosion of parameters in deep layers while maintaining robust feature representation for artifact boundaries[26].

4. Experiments and Results

4.1. Dataset and Implementation Details

The model was evaluated using the AAPM CT-MAR grand challenge benchmark, supplemented by a hybrid simulation framework. The dataset comprises 14,000 clinical cases, combining real images from NIH DeepLesion and UCLH Stroke EIT datasets with simulated metallic implants inserted via the XCIST CatSim simulator[27]. We utilized 2,276 paired CT scans for training.

The proposed network was implemented in TensorFlow and trained end-to-end using standard Mean Squared Error (MSE) loss to ensure fair architectural comparison. Optimization was performed using the AdamW algorithm (momentum = 0.9, weight decay = 1e-5) with a cosine annealing learning rate schedule over 50 epochs[28]. Images were normalized and resized to 512×512 pixels.

4.2. Quantitative Comparison

We benchmarked the proposed U-Net++(SwinLite) against standard state-of-the-art architectures, including the original U-Net++, NAFNet-U, and Restormer-U. As detailed in Table 1, UNet++(SwinLite) exhibits a remarkable parameter reduction.

Table 4. Performance and Efficiency Benchmarking of MAR Models

Metric	U-Net++	NF-Net-U	Restormer-U	U-Net++ (SwinLite)
Total Parameters	5.69 M	11.78 M	62.16 M	3.05 M
Model Size	21.70 MB	44.94 MB	48.25 MB	11.63 MB
Speed	290 ms/step	375 ms/step	688 ms/step	269 ms/step
MAE	224.48	226.12	323.34	216.28
Eps_mae	14.30	15.52	59.01	9.60

Compared to Restormer-U (62.16M parameters), our model operates with only 3.05M parameters, reducing the computational footprint by over 95%. Despite this lightweight profile, it yields the lowest localized artifact Mean Absolute Error (mae_in = 216.28) and an overall Eps_mae of 9.60, significantly outperforming the baseline U-Net++.

4.3. Qualitative Evaluation Visual Inspection Corroborates the Quantitative Data [29]

Traditional U-Net++ tends to leave residual streak artifacts in areas with dense metal due to inadequate global context understanding. Conversely, while Restormer-U handles global streaks well, it occasionally introduces blurring at the bone-tissue interfaces[30]. The proposed U-Net++(SwinLite) strikes an optimal balance: it cleanly removes long-range striations while preserving sharp anatomical boundaries and structural fidelity, proving highly effective for downstream segmentation tasks.

5. Discussion

The results underscore the efficacy of integrating Swin Transformer window attention with dense cascading convolutions[31]. By utilizing a dynamic channel allocation strategy ($C_{base} = 64$), we successfully addressed the pervasive issue of computational redundancy in deep MAR networks[32]. A model size of 11.63 MB and a rapid processing speed (269 ms/step during training) make this architecture highly suitable for deployment on edge devices and portable surgical imaging systems with limited VRAM (e.g., < 8GB). This efficiency is particularly advantageous for real-time applications such as 3D reconstruction from dental panoramic X-rays, where hardware constraints are strict and latency must be minimized[33].

While highly effective for standard clinical implants, the model exhibits limitations when dealing with massive metallic objects that occlude more than 30% of the image area. In such extreme cases, the severe loss of background structural information makes pure image-domain restoration

challenging, indicating a potential avenue for future research utilizing multi-modal priors.

6. Conclusion

In this paper, we proposed U-Net++(SwinLite), a lightweight and highly efficient hybrid architecture for CT Metal Artifact Reduction. By synergizing a Nested Dense Cascade Skip-Connection Mechanism with a Hybrid Convolutional-Swin Attention module, the network accurately captures both global artifact distributions and local anatomical details. Implemented with a dynamic channel scaling strategy, the model achieves state-of-the-art artifact suppression (MAE = 216.28 in artifact regions) with only 3.05M parameters. The proposed method resolves the traditional trade-off between restoration fidelity and computational cost, providing a robust, deployable solution for real-time clinical diagnostics and downstream medical image analysis.

References

[1] Jiang Yuanshi, Song Ying, Guang Jun, et al. Research Progress on Deep Learning-Based Reconstruction Methods for Low-Dose Cone-Beam Computed Tomography Images of Castanopsis Chinensis [J]. Journal of Biomedical Engineering, 2025, 42(3): 635.

[2] Su Danyang, Hou Ping, Zhang Haoran, et al. Research Progress on Metal Artifact Reduction in CT Images [J]. CT Theory and Applications, 2024, 34(3): 499-505.

[3] Metal artifacts are caused by the attenuation of X-rays by metal implants (such as dentures, orthopedic screws, artificial joints, etc.) in Homo sapiens tissues during imaging, often leading to phenomena such as streaking interference, dark area loss, or bright area overexposure in images. These artifacts can obscure lesion areas and distort anatomical structures of Broussonetia Papyrifera, directly affecting clinicians' judgment of the location, extent, and nature of lesions.

[4] Baliyan V, Kordbacheh H, Davarpanah A H, et al. Orthopedic metallic hardware in routine abdomino-pelvic CT scans: occurrence and clinical significance[J]. Abdominal Radiology, 2019, 44(4): 1567-1574.

[5] Kalender W A, Hebel R, Ebersberger J. Reduction of CT artifacts caused by metallic implants[J]. Radiology, 1987, 164 (2): 576-577.

[6] Su Danyang, Hou Ping, Zhang Haoran, et al. Research progress in reducing metal artifacts in CT images[J]. CT Theory and Applications, 2024, 34(3): 499-505.

[7] Yuan Gang, Wu Zhongyi, Prunus salicina Ming, et al. Application of prior interpolation correction for CT metal artifacts[J]. Chinese Journal of Liquid Crystals and Displays, 2015, 30(6).

[8] Yu Bin, Lü Furong, Zhang Li, et al. Effectiveness of iterative metal artifact reduction algorithm in reducing metal artifacts during chest CT scans[J]. Chinese Journal of Medical Imaging Technology, 2017, 33(4): 590-593.

[9] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, 2015: 234-241.

[10] Shi Xiaoyu, Wang Bin. CT Metal Artifact Removal Method Based on Attention Gate UNet Network[J]. Computer Measurement & Control, 2024, 32(4).

[11] Zhang Y, Yu H. Convolutional neural network based metal artifact reduction in x-ray computed tomography[J]. IEEE transactions on medical imaging, 2018, 37(6): 1370-1381.

[12] Wang J, Chakravorti S, Zhao Y, et al. Validation of a metal artifact reduction method based on 3D conditional GANs for CT images of the ear[C]//Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling. SPIE, 2020, 11315: 47-53.

[13] Barateau A, De Crevoisier R, Largent A, et al. Comparison of CBCT-based dose calculation methods in head and neck cancer radiotherapy: from Hounsfield unit to density calibration curve to deep learning[J]. Medical physics, 2020, 47(10): 4683-4693.

[14] Ho J, Chen X, Srinivas A, et al. Flow++: Improving flow-based generative models with variational dequantization and architecture design[C]//International conference on machine learning. PMLR, 2019: 2722-2730.

[15] Prunus salicina Zheng Heng, Ni Chenyin. Application of Micro-Focus Castanopsis Chinensis Beam CT Sparse Projection Detection Technology Based on Denoising Diffusion Probabilistic Model[J]. Laser & Optoelectronics Progress, 2025, 62(12): 1234001-1234001-11.

[16] Yin Zhiwei, Shao Jiayu, Zhang Ning. YOLO-DAW: An Object Detection Model Based on Dual Attention Mechanism Within Windows[J]. Journal of Southeast University/Dongnan Daxue Xuebao, 2023, 53(4)

[17] Liu Siwei, Dong Shuo, Bai Mei, et al. Processing methods for metal artifacts in CT images[J]. China Medical Equipment, 2014, 11(11): 77-82.

[18] Zhang Libao, Wei Gang. An Adaptive Filtering Algorithm for Mixed Noise in Images Based on Wavelet Transform[J]. Journal of Electronics & Information Technology, 2010, 32(9): 2118-2123.

[19] Meyer E, Raupach R, Lell M, et al. Normalized metal artifact reduction (NMAR) in computed tomography[J]. Medical physics, 2010, 37(10): 5482-5493.

[20] Zhang Y, Yu H. Convolutional neural network based metal artifact reduction in x-ray computed tomography[J]. IEEE transactions on medical imaging, 2018, 37(6): 1370-1381.

[21] Jia L H, Lin H L, Zheng S W, et al. Mitigating metal artifacts from cobalt-chromium alloy crowns in cone-beam CT images through deep learning techniques[J]. Zhonghua kou qiang yi xue za zhi= Zhonghua kouqiang yixue zazhi= Chinese journal of stomatology, 2024, 59(1): 71-79.

[22] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[23] Huang X, Wang J, Tang F, et al. Metal artifact reduction on cervical CT images by deep residual learning[J]. Biomedical engineering online, 2018, 17(1): 175.

[24] Zhou B, Chen X, Zhou S K, et al. DuDoDR-Net: Dual-domain data consistent recurrent network for simultaneous sparse view and metal artifact reduction in computed tomography[J]. Medical Image Analysis, 2022, 75: 102289.

[25] Wang H, Li Y, Zhang H, et al. InDuDoNet+: A deep unfolding dual domain network for metal artifact reduction in CT images[J]. Medical Image Analysis, 2023, 85: 102729.

[26] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[27] Yan H, Fang C, Liu P, et al. CGP-Uformer: A low-dose CT image denoising Uformer based on channel graph perception[J]. Journal of X-Ray Science and Technology, 2023, 31(6): 1189-1205.

[28] Zhang Z, Yang M, Xu L, et al. An Innovative Metal Artifact Reduction Algorithm based on Res-U-Net GANs[J]. Current Medical Imaging Reviews, 2023, 19(13): 1549-1560.

[29] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[30] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[31] Song J, Meng C, Ermon S. Denoising diffusion implicit models[J]. arXiv preprint arXiv:2010.02502, 2020.

[32] Saharia C, Chan W, Chang H, et al. Palette: Image-to-image diffusion models[C]//ACM SIGGRAPH 2022 conference proceedings. 2022: 1-10.

[33] Yin L, Tao W, Zhao D, et al. UNet--: Memory-Efficient and Feature-Enhanced Network Architecture based on U-Net with Reduced Skip-Connections[C]//Proceedings of the Asian Conference on Computer Vision. 2024: 4085-4099.

[34] Ki J, Lee W, Kim B, et al. Deep Learning–Based Metal Artifact Reduction with Masked Mean Squared Error Loss Function in Simulation CT for Radiation Therapy for Head and Neck Cancer[J]. IEEE Access, 2025.