RISE-MAR: Radiologist-Integrated Self-Evolution for Generalizable Metal Artifact Reduction in CT Imaging

Saud Muhammad Alshammari[1], Alaa Saud Alanazi[2], Hamdan Ahmed Hamdan Alshehri[3], Asmaa Alahmadi[4], Amer Altarjami[5], Nasser Almohsen[6], Nawaf Othman Ahmed Alyahyawi[7], Abdulaziz Muhammad alharbi[8], Muhannad Alahmade[9], Amjad Majed Alharbi[10], Hatun Eid Albishi[11], Ahmed Theyab Alsahli[12], Abdullah Mahmoud Alsenani[13], Abdulrahman Bassam Alhazmi[14], Ammar Hani Alhejaili[15]

[1] Radiology Technologist, Saudi Arabia,
[2] Radiology Technologist, Saudi Arabia,
[3] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[4] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[5] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[6] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[7] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[8] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[9] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[10] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[11] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[12] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[13] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[14] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia
[15] Radiology Technologist, Ministry of National Guard, Prince Mohammed bin Abdulaziz Hospital, Saudi Arabia

Accepted: 15-09-2025 Published: 15-10-2025

Abstract

This study introduces RISE-MAR (Radiologist-Integrated Self-Evolution for Metal Artifact Reduction), a novel framework that addresses the persistent challenge of metal artifacts in computed tomography (CT) imaging. Unlike previous approaches that rely solely on mathematical optimization or deep learning, RISE-MAR explicitly integrates radiologist expertise into a self-evolving system through an innovative feedback loop. Our dual-domain architecture combines a transformer-based image branch with a specialized sinogram network, enforcing physical consistency while capturing long-range dependencies essential for complex artifact patterns. The key innovation lies in our confidence-guided pseudo-labeling mechanism that selectively identifies high-confidence regions for self-training, refined by radiologist feedback that ensures clinical relevance. Experimental results demonstrate RISE-MAR's superior performance across synthetic and clinical datasets, with significant improvements in both quantitative metrics (PSNR: 36.2 dB, SSIM: 0.923) and clinical relevance scores (4.2/5) compared to state-of-the-art methods. Most notably, RISE-MAR shows remarkable generalization capability across different implant types and anatomical regions, effectively bridging the domain gap between training environments and clinical applications. Our work establishes a new paradigm for developing medical image enhancement techniques that leverage the complementary strengths of computational methods and clinical expertise.

1. INTRODUCTION: THE INTERPRETIVE CHALLENGE OF METAL ARTIFACTS

In the landscape of medical imaging interpretation, metal artifacts in Computed Tomography (CT) represent a particularly vexing hermeneutic challenge. These visual disruptions—streaking, shadowing, and distortion patterns that emanate from metallic implants—create a complex palimpsest where the original anatomical narrative becomes obscured by technical interference (Gjesteby et al., 2016). The resulting image resembles a contested text where multiple competing narratives vie for the reader's attention, complicating the diagnostic process and potentially leading to misinterpretation with serious clinical consequences.

The presence of these artifacts transforms the straightforward clinical reading into what Umberto Eco might call an "open work"—a text whose interpretation is radically unstable and subject to multiple competing readings. In radiation therapy planning, for instance, these artifacts can lead to significant errors in electron density calculations, compromising treatment accuracy (Kilby et al., 2002). The metallic objects—hip prostheses, dental implants, spinal fixation devices, and other surgical hardware—create a discursive rupture in the CT narrative, introducing what Roland Barthes might term "noise" that disrupts the signifying process.

The history of metal artifact reduction (MAR) presents a fascinating parallel to the evolution of literary critical methodologies over the past century. The earliest approaches, pioneered by Kalender et al. (1987), employed a form of technical close reading focused on sinogram manipulation—identifying the metal trace, excising it from the narrative, and reconstructing the missing information through interpolation. This resembles New Criticism's focus on the text itself, bracketing external context in favor of internal coherence. While groundbreaking, these approaches often generated secondary artifacts—new interpretive problems that arose from the very attempt to restore the original image.

As MAR techniques evolved, they began to incorporate more contextual information, similar to how literary criticism moved beyond pure formalism. Meyer et al. (2010) introduced Normalized Metal Artifact Reduction (NMAR), which acknowledged the importance of tissue classification—a form of contextual knowledge—in preventing secondary artifacts. This represented a shift toward a more holistic reading that considered the relationship between elements rather than treating them in isolation.

The emergence of deep learning approaches in MAR parallels the post-structuralist turn in literary theory, with its emphasis on complex systems of signification and the instability of meaning. Zhang and Yu (2018) demonstrated that convolutional neural networks could learn to map artifact-corrupted images to clean ones through supervised training—essentially identifying deep structural patterns beneath surface distortions. This computational approach to pattern recognition resembles the structuralist search for underlying codes and conventions that generate meaning.

More recent dual-domain approaches like DuDoNet (Lin et al., 2019) and Quad-Net (Li et al., 2024) acknowledge the dialogic relationship between different representational domains—image and sinogram—recognizing that meaning emerges through their interplay. This parallels how post-structuralism emphasizes the relationships between texts and the ways meaning arises through difference rather than inherent properties. Despite these advances, a significant interpretive gap remains between training environments and clinical applications (Du et al., 2021), resembling the tension between theoretical frameworks and practical criticism.

In this study, we introduce RISE-MAR (Radiologist-Integrated Self-Evolution for Metal Artifact Reduction), a framework that draws inspiration from reader-response criticism and the concept of interpretive communities developed by Stanley Fish. Our approach acknowledges that the interpretation of medical images, like literary texts, cannot be divorced from the communities that read them. By integrating radiologist expertise into a self-evolving MAR system, we create a dialectical process that bridges the gap between computational analysis and clinical judgment, producing artifact reduction that is both technically accurate and clinically relevant.

2. The Evolving Canon of Artifact Reduction: A Critical Genealogy

2.1 Formalist Beginnings: Close Reading the Sinogram

The earliest MAR techniques emerged in the late 1980s, when Kalender et al. (1987) introduced what would become the foundational approach to metal artifact reduction. Their method, akin to New Critical close reading, focused intensely on the sinogram—the Radon transform of the CT image—as the primary text to be interrogated. The approach identified metal traces within this domain, excised them through segmentation, and then employed linear interpolation to reconstruct the missing data.

This formalist approach, with its emphasis on the internal coherence of the sinogram, achieved notable success in reducing primary streaking artifacts. However, much like New Criticism's struggle with contextual elements that resist purely formal analysis, this method often introduced secondary artifacts—new visual distortions that emerged from the interpolation process itself. The interpolation, while mathematically elegant, failed to account for the complex anatomical variations that characterize human tissue, resulting in a flattened representation that lacked textural nuance.

Subsequent analytical approaches attempted to refine this formalist reading through more sophisticated interpolation techniques. Kalender's original linear interpolation gave way to higher-order polynomial methods, spline-based approaches, and eventually directional interpolation that considered the orientation of structures near the metal. Each refinement represented an attempt to perform a more nuanced close reading of the sinogram, acknowledging its internal complexities while still maintaining the formalist emphasis on the text itself.

Meyer et al. (2010) made a significant advance with Normalized Metal Artifact Reduction (NMAR), which incorporated tissue classification into the interpolation process. This represented a shift away from pure formalism toward a more contextually informed reading. By normalizing the sinogram based on a prior image derived from tissue segmentation, NMAR acknowledged that the meaning of specific sinogram values depends on their relationship to surrounding tissues—a kind of contextual significance that pure formalism might overlook.

Despite these advances, traditional analytical approaches remained limited by their inability to capture the full complexity of metal artifacts, particularly in cases involving multiple or large metallic implants. Much as formalist criticism struggled with highly allusive or contextually embedded texts, these methods often failed when confronted with complex artifact patterns that required broader contextual understanding for proper interpretation.

2.2 The Structuralist Turn: Deep Learning and Pattern Recognition

The application of deep learning to MAR represented a paradigm shift akin to the structuralist turn in literary criticism. Rather than focusing on explicit rules for artifact correction, these approaches sought to identify underlying patterns and transformational principles through data-driven learning. Zhang and Yu (2018) demonstrated the potential of this approach by training convolutional neural networks to map artifact-corrupted images directly to clean ones, bypassing the explicit modeling of artifact formation.

This structuralist-like approach treated the artifact reduction problem as one of pattern recognition and transformation, similar to how structuralist critics sought to identify the underlying codes and conventions that generate meaning in texts. The CNN architecture, with its hierarchical feature extraction capabilities, functioned as a kind of computational structuralist, identifying patterns at multiple scales and learning the transformational rules that could convert corrupted images into clean ones.

The supervised learning paradigm employed in early deep learning MAR approaches required paired datasets—artifact-free and artifact-corrupted versions of the same images. This parallels how structuralist criticism often relied on comparative analysis of multiple texts to identify common patterns and variations. However, this requirement presented a significant limitation in clinical settings, where obtaining perfectly matched artifact-free and artifact-corrupted images is impossible for the same patient.

A notable limitation of these image-domain approaches was their disconnection from the physical process of CT image formation. By operating solely in the image domain, they resembled a form of criticism that focuses exclusively on the final text while ignoring its production process. This limitation led to the development of more sophisticated approaches that acknowledged the multi-domain nature of CT imaging.

2.3 Post-Structuralist Complexity: Multi-Domain Approaches

The development of dual-domain and multi-domain MAR approaches parallels post-structuralism's emphasis on intertextuality and the complex relationships between signifying systems. DuDoNet (Lin et al., 2019) pioneered this approach by explicitly modeling the relationship between image and sinogram domains through a consistency layer that enforced the Radon transform relationship between them.

This dual-domain approach acknowledged that meaning in CT imaging emerges through the dialogue between different representational systems—the image domain, where radiologists perform their readings, and the sinogram domain, where the physical process of CT imaging occurs. The consistency layer functioned as a kind of intertextual constraint, ensuring that interpretations in one domain respected the physical realities represented in the other.

Liao et al. (2019) further developed this multi-domain discourse with their Artifact Disentanglement Network (ADN), which sought to separate metal artifacts from anatomical structures through an unsupervised approach. This disentanglement strategy resembles deconstructionist criticism's attempt to identify and separate the competing discourses that constitute a text, revealing the tensions and contradictions within seemingly unified representations.

More recently, Li et al. (2024) introduced Quad-Net, which extends the multi-domain approach to include wavelet transforms alongside image and sinogram representations. This increasing complexity parallels how post-structuralist criticism embraces multiple theoretical frameworks simultaneously, acknowledging that different interpretive lenses reveal different aspects of textual meaning.

Lyu et al. (2021) contributed significantly to this discourse with U-DuDoNet, which eliminated the need for perfectly aligned artifact-free and artifact-corrupted image pairs through an unpaired training approach. This innovation parallels how post-structuralism challenged the structuralist assumption of stable binary oppositions, embracing instead a more fluid conception of textual relationships.

2.4 The Contextual Challenge: Domain Adaptation and Transfer

Despite these advances in MAR methodology, a significant challenge remained: the domain gap between training data and clinical applications. Models trained on simulated data or specific metal artifact patterns often failed when confronted with the heteroglossia of real clinical scenarios—the diverse implant types, varying CT acquisition parameters, and patient-specific anatomical variations that characterize medical practice.

This challenge parallels the tension between theoretical frameworks and practical criticism in literary studies—the gap between abstract models and the messy reality of specific texts in particular contexts. Domain adaptation techniques emerged as an attempt to address this gap, similar to how contextualist approaches in literary criticism sought to acknowledge the situational embeddedness of texts.

Ganin et al. (2016) introduced Domain-Adversarial Neural Networks (DANN), which attempt to learn domain-invariant features through adversarial training. This approach resembles archetypal criticism's search for transcendent patterns that persist across different cultural and historical contexts. By forcing the feature extractor to produce representations that cannot be distinguished by domain, DANN attempts to identify a kind of universal visual language that transcends specific imaging contexts.

Du et al. (2023) applied unsupervised domain adaptation to MAR, demonstrating improved performance on clinical data after training on synthetic datasets. Their approach acknowledged the contextual nature of artifact interpretation while seeking to bridge different contextual frameworks through adaptive regularization techniques.

Wang et al. (2023) introduced SemiMAR, a semi-supervised learning approach that leverages a teacher-student framework to generate pseudo-labels for unlabeled data. This approach resembles how emerging critics learn interpretive strategies from established scholars, gradually developing their own readings while being guided by authoritative interpretations. However, SemiMAR still lacked mechanisms to ensure the clinical relevance of its interpretations, potentially suffering from confirmation bias as the model reinforced its own misreadings through erroneous pseudo-labels.

3. RISE-MAR: A Reader-Response Framework for Image Interpretation

3.1 Theoretical Foundations: The Interpretive Community

Our RISE-MAR framework draws theoretical inspiration from reader-response criticism, particularly Stanley Fish's concept of interpretive communities. Fish argued that meaning does not reside in texts themselves but emerges through the interpretive strategies employed by communities of readers who share certain assumptions, knowledge, and reading practices. In the context of CT imaging, radiologists constitute such an interpretive community, bringing specialized knowledge and clinical priorities to their reading of medical images.

The conventional approach to MAR development has emphasized technical metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), which measure mathematical fidelity to ground truth images. This parallels how formalist criticism might emphasize technical elements like meter, rhyme, and figurative language while potentially overlooking the meaning that emerges through reader engagement. Our approach acknowledges that clinical relevance—the usefulness of an image for diagnostic purposes—cannot be reduced to these technical metrics alone.

RISE-MAR explicitly incorporates the radiologist's interpretive expertise through a feedback loop that guides the evolution of the artifact reduction system. This approach recognizes that the ultimate goal of MAR is not perfect mathematical reconstruction but clinically useful images that support accurate diagnosis and treatment planning. By integrating radiologist feedback, we ensure that the system learns to prioritize the aspects of image quality that matter most in clinical practice.

3.2 Architectural Components: A Dialectical Approach

The RISE-MAR framework consists of four main components that work together in a dialectical process of interpretation and refinement:

1. **Dual-Domain Network**: This component performs the initial reading of the artifact-corrupted image, operating across both image and sinogram domains to generate a provisional artifact-reduced interpretation.
2. **Confidence-Guided Pseudo-Labeling**: This mechanism selectively identifies high-confidence interpretations from unlabeled data, addressing the problem of confirmation bias in self-training.
3. **Radiologist Feedback Integration**: This module incorporates expert judgment into the self-training loop, ensuring that the model evolution aligns with clinical priorities.
4. **Model Evolution Strategy**: This component manages the dialectical process through which the model improves over time, balancing different learning objectives and preventing catastrophic forgetting.

3.2.1 Dual-Domain Network: Close Reading Across Representational Levels

Our dual-domain network builds upon previous approaches but introduces several key innovations designed to enhance its interpretive capabilities. In the image domain branch, we implement a transformer-based architecture inspired by the Swin Transformer (Liu et al., 2021), which employs shifted windows to efficiently model hierarchical features.

This architectural choice parallels how sophisticated literary analysis attends to both local textual details and broader structural patterns. The self-attention mechanism in transformers allows the model to consider long-range dependencies between image regions, similar to how critics might trace motifs or themes across an entire text. The hierarchical structure of the Swin Transformer, with its progressive merging of attention windows, resembles the way critics move between close analysis of specific passages and consideration of larger narrative or thematic structures.

The sinogram domain branch employs a U-Net architecture specialized for sinogram completion, with skip connections that preserve detailed information across different resolution levels. This branch performs a kind of close reading of the raw CT data, attending to the physical process through which the image is formed rather than just its final appearance.

The domain transformation module implements the Radon and inverse Radon transforms, enabling bidirectional conversion between image and sinogram domains. This module enforces physical consistency between the two domains through a specialized loss function:
$\mathcal{L}{consistency} = ||R(\hat{x}{clean}) - \hat{s}{clean}||1 + ||R^{-1}(\hat{s}{clean}) - \hat{x}{clean}||_1$

This consistency constraint functions as a form of intertextual coherence, ensuring that interpretations in one domain respect the physical realities represented in the other. It parallels how literary critics might check their interpretations of a primary text against related documents, historical contexts, or authorial statements to ensure coherence across different textual levels.

3.2.2 Confidence-Guided Pseudo-Labeling: Selective Attention and Interpretive Uncertainty

A critical challenge in self-training approaches is confirmation bias—the tendency for models to reinforce their own errors through erroneous pseudo-labels (Arazo et al., 2020). This parallels how readers may selectively attend to textual elements that confirm their preexisting interpretations while overlooking contradictory evidence.

Our confidence-guided pseudo-labeling mechanism addresses this challenge by incorporating a measure of interpretive uncertainty into the self-training process. For each unlabeled image x_u, we compute artifact-reduced outputs from both the image and sinogram branches:

$\hat{x}_u = G_I(x_u)$

$\hat{s}_u = G_S(R(x_u))$

$\hat{x}_u' = R^{-1}(\hat{s}_u)$

We then calculate a confidence score based on the consistency between these interpretations:

$c_u = \exp(-||R(\hat{x}_u) - \hat{s}_u||_1 - ||R^{-1}(\hat{s}_u) - \hat{x}_u||_1)$

This score reflects the degree of agreement between different "readings" of the same data. High consistency suggests a more reliable interpretation, while discrepancies indicate interpretive uncertainty. We select pseudo-labels for which $c_u > \tau_t$, where τ_t is a dynamic threshold that increases as the model evolves.

This selective approach to pseudo-labeling parallels how readers develop metacognitive awareness of their own interpretive certainty, giving more weight to readings they can justify through multiple textual warrants and approaching uncertain interpretations with appropriate caution. The dynamic threshold reflects how interpretive standards become more rigorous as readers develop greater expertise, demanding stronger evidence for claims as their critical faculties mature.

To further enhance the reliability of pseudo-labels, we apply a weighted combination of the image and sinogram domain outputs:

$\hat{x}_{pseudo} = \lambda_t \cdot \hat{x}_u + (1-\lambda_t) \cdot \hat{x}_u'$

This weighted approach resembles how readers might synthesize different interpretive frameworks, giving more weight to approaches that have proven more reliable in specific contexts while still incorporating insights from multiple perspectives.

3.2.3 Radiologist Feedback Integration: The Expert Reader in the Loop

The most distinctive feature of our RISE-MAR framework is the integration of expert radiologist feedback into the self-training loop. After generating and filtering pseudo-labels through our confidence-guided mechanism, a subset of these pseudo-labels is presented to radiologists for review.

Radiologists assess the clinical relevance of the artifact reduction by comparing the original artifact-corrupted image with the pseudo-label, focusing on three criteria:

1. **Effective artifact reduction**: Has the MAR process successfully reduced visually distracting artifacts?
2. **Preservation of anatomical structures**: Are clinically important anatomical details preserved without blurring or distortion?
3. **Absence of new artifacts or hallucinations**: Has the process avoided introducing new visual anomalies that might be misinterpreted as pathology?

Based on these criteria, radiologists assign a clinical relevance score (0-5) to each reviewed pseudo-label. This score represents the expert reader's assessment of the interpretation's usefulness within the clinical interpretive community.

These scores influence the learning process in two crucial ways:

First, only pseudo-labels with scores above a threshold are included in the training set. This reflects how interpretive communities validate certain readings while rejecting others based on shared standards of evidence and relevance.

Second, the scores are used to weight the contribution of each pseudo-label to the training loss:

$\mathcal{L}{pseudo} = \sum{i=1}^{N_p} w_i \cdot ||G(x_i) – \hat{x}_{pseudo,i}||_1$

where w_i is derived from the radiologist's score. This weighted approach ensures that the model prioritizes learning from clinically valuable examples, similar to how literary scholars might give more weight to certain exemplary readings that demonstrate particularly insightful applications of critical principles.

To use radiologists' time efficiently, we employ an active learning strategy that selects the most informative examples for review. This strategy prioritizes examples with moderate confidence scores—where the model shows some uncertainty but not complete confusion—and examples representing diverse metal artifact patterns and anatomical regions. This selective approach parallels how teachers of literary criticism might choose particularly instructive examples for close analysis rather than attempting comprehensive coverage of every possible text.

3.2.4 Model Evolution: The Dialectical Development of Interpretation

Our model evolution strategy combines multiple learning objectives in a dialectical process that gradually improves the system's interpretive capabilities. The overall training objective consists of several components:

$\mathcal{L}{total} = \mathcal{L}{supervised} + \alpha_t \cdot \mathcal{L}{pseudo} + \beta_t \cdot \mathcal{L}{consistency} + \gamma_t \cdot \mathcal{L}_{adversarial}$

The supervised loss $\mathcal{L}_{supervised}$ is computed on the labeled synthetic dataset, where ground truth artifact-free images are available:

$\mathcal{L}{supervised} = ||G_I(x{art}) - x_{clean}||1 + ||G_S(s{art}) - s_{clean}||_1$

This resembles how literary critics ground their interpretations in close attention to textual details, establishing a foundation of textual evidence before developing more speculative readings.

The pseudo-label loss $\mathcal{L}_{pseudo}$ incorporates radiologist-approved interpretations from unlabeled clinical data, representing the integration of expert judgment with formal analysis.

The consistency loss $\mathcal{L}_{consistency}$ enforces coherence between image and sinogram domain interpretations, reflecting the expectation that valid readings maintain coherence across different textual levels and contexts.

The adversarial loss $\mathcal{L}_{adversarial}$ employs a Wasserstein GAN with gradient penalty (Arjovsky et al., 2017), encouraging the generated artifact-reduced images to match the distribution of artifact-free images:

$\mathcal{L}{adversarial} = \mathbb{E}{x_{art}}[D(G_I(x_{art}))] - \mathbb{E}{x{clean}}[D(x_{clean})] + \lambda_{gp} \cdot \mathcal{L}_{gp}$

This adversarial component parallels how literary interpretation involves not just analysis of specific textual features but also consideration of how texts conform to or deviate from generic conventions and reader expectations.

The time-dependent weighting factors α_t, β_t, and γ_t balance these different objectives during the evolution process. Their dynamic adjustment reflects how interpretive priorities shift as critical understanding develops, with different aspects of textual analysis taking precedence at different stages of the interpretive process.

To prevent catastrophic forgetting during model evolution, we employ a knowledge distillation approach where the model retains a memory of previous iterations:
$\mathcal{L}{distill} = ||G_t(x) - G{t-1}(x)||_1$
This distillation loss ensures continuity with previous interpretive frameworks, reflecting how literary criticism typically builds upon rather than simply replaces earlier approaches. Even as new theoretical frameworks emerge, they often incorporate insights from previous critical traditions, maintaining a dialogue with earlier approaches even as they move beyond them.

4. Experimental Dialogue: Testing the Interpretive Framework

4.1 The Corpus: Diverse Textual Traditions

We evaluated our RISE-MAR framework on three distinct corpora representing different domains within the broader discourse of CT imaging:

1. **Synthetic Corpus**: Created by simulating metal artifacts on the DeepLesion dataset (Yan et al., 2018), this corpus of 32,735 CT slices from 4,427 patients provides what might be considered "critical editions" where both the artifact-corrupted and artifact-free versions are available for comparison. The metal artifacts were simulated by inserting virtual metallic implants of various shapes and sizes, followed by forward and backward projections to generate physically realistic artifacts. This corpus was divided into training (80%), validation (10%), and testing (10%) sets.
2. **Clinical Corpus A**: This collection of 1,200 CT slices from 150 patients with hip implants represents authentic clinical texts where no artifact-free "original" exists. Collected from a single institution, these images present the real-world complexity of metal artifacts in orthopedic imaging, with variations in implant type, size, and orientation. This corpus was divided into unlabeled training (70%) and testing (30%) sets.
3. **Clinical Corpus B**: To test cross-institutional and cross-anatomical interpretation, we assembled 800 CT slices from 100 patients with dental implants from a different institution. This corpus represents a related but distinct textual tradition, with different artifact patterns, image acquisition parameters, and anatomical contexts. It was used exclusively for testing, assessing the generalization capability of our interpretive framework.

For the radiologist feedback integration, we recruited five board-certified radiologists with specialization in musculoskeletal and head and neck imaging, representing the expert readers within these interpretive communities. Their experience ranged from 5 to 15 years of post-fellowship practice, providing a range of expertise levels within the professional community. Throughout the model evolution process, each radiologist reviewed approximately 200 pseudo-labeled examples, providing clinical relevance scores and qualitative feedback.

4.2 Implementation Details: The Mechanics of Interpretation

Our dual-domain MAR network was implemented using PyTorch, with the image domain branch consisting of a Swin Transformer with four stages and embedding dimensions of [96, 192, 384, 768]. The sinogram domain branch employed a U-Net architecture with five encoder-decoder layers and skip connections. The domain transformation module implemented the Radon and inverse Radon transforms using the ASTRA toolbox, providing physically accurate conversion between domains.

For model training, we used the Adam optimizer with an initial learning rate of 1e-4, β1 = 0.5, β2 = 0.999, and a batch size of 16. A cosine learning rate schedule with warm-up was employed to stabilize early training. The weighting factors α_t, β_t, and γ_t were initialized to 0.1, 1.0, and 0.01 respectively, and adjusted dynamically based on validation performance using a grid search strategy.

The confidence threshold τ_t for pseudo-label selection was initialized to 0.8 and increased by 0.02 in each iteration until reaching 0.9, reflecting the increasing selectivity of the pseudo-labeling process as the model evolved. The weighting factor λ_t for combining image and sinogram domain outputs was set to 0.7 initially and adjusted based on the relative performance of the two branches on the validation set.

The model evolution process consisted of 10 iterations, with each iteration involving pseudo-label generation, radiologist review, and model update. In each iteration, approximately 500 high-confidence pseudo-labels were generated, from which 100-200 were selected for radiologist review using our active learning strategy. This selective approach made efficient use of radiologist time while still providing sufficient expert guidance for model evolution.

4.3 Comparative Methods: Alternative Interpretive Frameworks

We compared RISE-MAR with several state-of-the-art MAR methods representing different approaches to the artifact reduction problem:

1. **Traditional Methods**:

o Normalized Metal Artifact Reduction (NMAR) (Meyer et al., 2010): A sophisticated analytical approach that incorporates tissue classification to prevent secondary artifacts.

2. **Deep Learning Methods**:

o CNN-MAR (Zhang & Yu, 2018): A CNN-based approach operating exclusively in the image domain, trained with supervised learning.

o DuDoNet (Lin et al., 2019): A dual-domain network that incorporates a sinogram consistency layer to enforce physical constraints.

o ADN (Liao et al., 2019): An artifact disentanglement network that separates metal artifacts from anatomical structures through unsupervised learning.

o Quad-Net (Li et al., 2024): A quad-domain network that incorporates wavelet transforms alongside image and sinogram domains.

3. **Domain Adaptation Methods**:

o DANN-MAR: An adaptation of Domain-Adversarial Neural Networks (Ganin et al., 2016) to the MAR task, incorporating adversarial domain adaptation.

o SemiMAR (Wang et al., 2023): A semi-supervised learning approach using a teacher-student framework to generate pseudo-labels for unlabeled data.

All methods were implemented following the descriptions in their respective papers, with hyperparameters optimized for best performance on our validation set. For fair comparison, all deep learning methods were trained using the same synthetic dataset and evaluated on the same test sets.

4.4 Evaluation Metrics: Quantifying Interpretive Quality

We employed both quantitative metrics and qualitative assessments to evaluate the performance of the MAR methods:

Quantitative Metrics:

- Peak Signal-to-Noise Ratio (PSNR): Measures the pixel-wise accuracy of the reconstructed image compared to the ground truth (when available).
- Structural Similarity Index (SSIM): Assesses the preservation of structural information and perceptual quality.
- Artifact Index (AI): A custom metric that quantifies the severity of streak artifacts by measuring the standard deviation in homogeneous regions near metal implants.
- Clinical Relevance Score (CRS): The average of radiologists' ratings on a 0-5 scale, reflecting the clinical utility of the artifact-reduced images.

Qualitative Assessment:

- Radiologist preference study: Pairwise comparisons between RISE-MAR and other methods, with radiologists indicating their preferred image for diagnostic purposes.

- Diagnostic confidence ratings: Radiologists' assessment of their confidence in making diagnoses from the artifact-reduced images, rated on a 1-5 scale.
- Structured evaluation of specific criteria: Effective artifact reduction, preservation of anatomical structures, and absence of new artifacts or hallucinations.

For the synthetic dataset, where ground truth artifact-free images were available, we could compute PSNR and SSIM directly. For the clinical datasets, where no ground truth exists, we relied more heavily on the Artifact Index, Clinical Relevance Score, and qualitative assessments.

5. RESULTS AND CRITICAL ANALYSIS

5.1 Quantitative Performance: Measuring Interpretive Fidelity

On the synthetic dataset (Table 1), where ground truth artifact-free images were available for comparison, RISE-MAR achieved superior performance across all quantitative metrics. The PSNR of 36.2 dB and SSIM of 0.923 represented improvements of 2.1 dB and 0.028, respectively, over the second-best method (Quad-Net). The Artifact Index showed a 43% reduction compared to Quad-Net, indicating substantially more effective streak artifact removal.

These quantitative improvements were most pronounced in cases involving complex metal artifacts from multiple or irregularly shaped implants—scenarios where the physical modeling of artifact formation becomes particularly challenging. In these cases, the radiologist guidance appeared to help the model prioritize the most clinically disruptive artifacts, leading to more perceptually relevant improvement even when overall mathematical fidelity (as measured by PSNR) showed more modest gains.

For Clinical Dataset A (hip implants), where no ground truth images were available, we relied on the Artifact Index and Clinical Relevance Score. RISE-MAR achieved an AI of 0.031, compared to 0.047 for Quad-Net and 0.052 for SemiMAR, indicating more effective reduction of streak artifacts in homogeneous tissue regions. More significantly, the CRS for RISE-MAR was 4.2 out of 5, substantially higher than SemiMAR (3.6) and Quad-Net (3.4).

This disparity between technical metrics and clinical relevance highlights the importance of expert judgment in evaluating MAR performance. While other methods achieved reasonable artifact reduction according to technical measures, radiologists consistently rated RISE-MAR's results as more clinically useful, suggesting that it better preserved diagnostically important features while removing distracting artifacts.

The most challenging evaluation came with Clinical Dataset B (dental implants), which represented both a different type of metal artifact and a different anatomical region from a different institution. Here, RISE-MAR maintained strong performance with a CRS of 3.9, while other methods showed significant degradation in performance (CRS ranging from 2.1 to 3.2).

This robust cross-domain performance demonstrates RISE-MAR's ability to generalize interpretive principles across different contexts—a key goal of our radiologist-integrated approach. The radiologist feedback appears to have guided the model toward more generalizable artifact reduction strategies that transfer effectively across different implant types and anatomical regions.

5.2 Ablation Study: Isolating Interpretive Components

To understand the contribution of each component of our RISE-MAR framework, we conducted a systematic ablation study (Table 4). This analysis helps isolate the specific impact of the radiologist feedback integration and confidence-guided pseudo-labeling mechanisms.

Removing the radiologist feedback integration while keeping other components intact resulted in a 1.8 dB decrease in PSNR on the synthetic dataset and a 0.4-point decrease in CRS on the clinical datasets. This substantial degradation highlights the critical role of expert guidance in shaping the model's learning trajectory. Without radiologist feedback, the model tended to prioritize artifact reduction based on mathematical criteria that did not always align with clinical relevance.

Replacing our confidence-guided pseudo-labeling with standard pseudo-labeling (using a fixed threshold) led to a 1.2 dB decrease in PSNR and a 0.3-point decrease in CRS. This confirms the importance of selective attention in the self-training process—by focusing on high-confidence predictions, our approach mitigates confirmation bias and prevents the reinforcement of errors through erroneous pseudo-labels.

Most interestingly, the combination of radiologist feedback and confidence-guided pseudo-labeling showed a synergistic effect. Removing both components simultaneously resulted in a 3.5 dB decrease in PSNR and a 0.9-point decrease in CRS—significantly greater than the sum of the individual component removals (3.0 dB and 0.7 points). This suggests that these components complement each other in important ways: radiologist feedback helps refine the confidence estimation mechanism, while the confidence-guided selection helps identify the most informative examples for radiologist review.

We also evaluated the impact of the transformer-based architecture by replacing it with a ResNet-based CNN while keeping other components intact. This resulted in a 1.1 dB decrease in PSNR and a 0.2-point decrease in CRS, confirming the advantage of the transformer's ability to capture long-range dependencies and hierarchical features in the artifact reduction task.

5.3 Qualitative Analysis: The Texture of Interpretation

Beyond quantitative metrics, the qualitative analysis of artifact-reduced images provides crucial insight into the interpretive texture achieved by different MAR methods. Figure 2 presents representative examples from all three datasets, comparing RISE-MAR with other approaches.

In the synthetic dataset examples, all methods showed some success in reducing primary streak artifacts. However, RISE-MAR demonstrated superior preservation of fine anatomical details near the metal implants—subtle tissue boundaries and small structures that other methods tended to blur or distort. This attention to fine detail parallels how sophisticated literary interpretation attends to nuance and ambiguity rather than imposing overly simplified readings.

The clinical dataset examples revealed more substantial differences between methods. In cases with hip implants (Clinical Dataset A), traditional methods like NMAR often introduced secondary artifacts—new visual distortions that replaced the original streaks. Deep learning methods generally avoided these secondary artifacts but sometimes produced oversmoothed regions near the implant boundaries, obscuring potentially important anatomical information.

RISE-MAR's results showed a more balanced approach, effectively removing distracting streaks while preserving tissue texture and anatomical boundaries. In several cases, subtle findings near implants—small bone fragments or soft tissue abnormalities—remained visible in RISE-MAR's output but were obscured in other methods' results. This preservation of diagnostically relevant details appears to be a direct result of the radiologist feedback integration, which guided the model to distinguish between artifacts that should be removed and anatomical variations that should be preserved.

The dental implant cases (Clinical Dataset B) presented even greater challenges due to the complex anatomy of the oral cavity and the smaller size but higher density of dental materials. Here, RISE-MAR demonstrated particularly strong generalization capability, effectively reducing artifacts while preserving the fine structures of teeth, the mandible, and surrounding soft tissues. Other methods either failed to adequately reduce artifacts or inappropriately smoothed anatomical structures, compromising diagnostic quality.

In the radiologist preference study, RISE-MAR was preferred in 78% of cases when compared to the second-best method (Quad-Net), with the remaining 22% rated as equivalent. None of the radiologists preferred other methods over RISE-MAR in direct comparison. When rating diagnostic confidence, radiologists reported an average increase of 1.7 points (on a 1-5 scale) when using RISE-MAR-processed images compared to artifact-corrupted originals, substantially higher than the increase for other methods (0.9-1.3 points).

5.4 Computational Considerations: The Economics of Interpretation

While RISE-MAR achieved superior performance across all evaluation metrics, it's important to consider the computational and resource implications of our approach. The radiologist feedback integration introduces a human-in-the-loop element that adds both value and cost to the MAR development process.

For the initial training and evolution of the RISE-MAR model, each participating radiologist spent approximately 10-15 hours reviewing pseudo-labeled examples and providing feedback over the course of the project. This represents a significant time investment from specialist professionals, though it should be noted that this investment is a one-time cost during model development rather than a recurring requirement during deployment.

Our active learning strategy helped maximize the efficiency of this radiologist time by selecting the most informative examples for review. We found that reviewing approximately 1,000 strategically selected examples over 10 iterations was sufficient to guide the model toward clinically relevant artifact reduction, a number that could be feasibly incorporated into existing workflow during model development.

In terms of computational resources, RISE-MAR's dual-domain transformer-based architecture requires more processing power than simpler CNN-based approaches. The inference time for a single CT slice was approximately 0.8 seconds on a NVIDIA A100 GPU, compared to 0.3-0.5 seconds for CNN-based methods. However, this remains well within the constraints of clinical workflow, where image processing typically occurs in parallel with other aspects of the radiological examination.

6. DISCUSSION: HERMENEUTIC IMPLICATIONS AND FUTURE DIRECTIONS

6.1 The Dialectic of Expertise and Automation

Our results demonstrate the value of integrating expert judgment with computational techniques in addressing complex interpretive challenges like metal artifact reduction. Neither purely algorithmic approaches nor unassisted human reading can achieve optimal results in isolation—it is through their dialogue that the most meaningful interpretations emerge.

This dialectical relationship between expertise and automation has broader implications for artificial intelligence in healthcare and beyond. Rather than positioning AI as a replacement for human expertise, our approach suggests a more productive framing in which computational methods amplify human capabilities while being guided by human judgment. The radiologist feedback in RISE-MAR doesn't simply validate the system's outputs but actively shapes its evolution toward clinically relevant performance.

This approach parallels developments in literary criticism, where computational methods like distant reading and topic modeling have emerged as supplements to traditional close reading rather than replacements for expert interpretation. In both domains, the most promising path forward appears to be one that recognizes the complementary strengths of human and machine intelligence.

6.2 The Politics of Interpretation: Who Decides What Matters?

An important aspect of our RISE-MAR framework is that it explicitly acknowledges the role of expert communities in determining interpretive priorities. By incorporating radiologist feedback into the self-training loop, we recognize that decisions about which artifacts matter most and which anatomical features must be preserved are not purely technical questions but matters of clinical judgment.

This raises broader questions about the politics of interpretation in AI systems. Who determines what constitutes a "good" interpretation? Whose priorities are encoded in the metrics and loss functions that guide model development? Our approach makes these questions explicit by directly incorporating expert feedback rather than relying solely on mathematical proxies for quality.

In the context of MAR, radiologists bring clinical priorities to their assessment—focusing on the preservation of diagnostically important features rather than mathematical fidelity to an idealized ground truth. This clinical perspective may sometimes diverge from purely technical metrics, as evidenced by the cases where RISE-MAR showed modest PSNR improvements but substantial gains in Clinical Relevance Score.

6.3 Limitations and Future Directions

Despite its strong performance, our approach has several limitations that suggest directions for future work. The requirement for radiologist feedback introduces a resource constraint, as expert time is limited and valuable. While our active learning strategy aims to minimize the number of examples requiring review, scaling to very large and diverse datasets remains challenging.

Future work could explore more efficient ways to incorporate expert feedback, such as through interactive tools that allow radiologists to quickly indicate regions of interest or concern within images. Sparse annotation approaches that focus radiologist attention on the most informative aspects of each image could further improve efficiency while maintaining the benefits of expert guidance.

Another limitation lies in the dual-domain architecture's requirement for access to sinogram data or the ability to perform accurate Radon transforms. In clinical settings with proprietary CT systems, direct access to raw sinogram data may be restricted. Future work could explore domain adaptation techniques to bridge the gap between idealized Radon transforms and the specific projection geometries of different CT manufacturers.

The transferability of radiologist feedback across different metal artifact types and anatomical regions also merits further investigation. While our results on Clinical Dataset B suggest good generalization capability, a more systematic exploration of how specific feedback translates across contexts could further enhance the framework's adaptability.

Finally, extending RISE-MAR to handle 3D volumes directly, rather than processing 2D slices independently, represents an important direction for future work. This volumetric approach could improve consistency across adjacent slices and better capture the three-dimensional nature of both anatomical structures and artifact patterns.

7. CONCLUSION: TOWARD A MORE NUANCED ARTIFACT REDUCTION

The RISE-MAR framework represents a synthesis of technical analysis and expert interpretation in addressing the challenge of metal artifacts in CT imaging. By acknowledging the essential role of the interpretive community in determining clinical relevance, our approach achieves more generalizable and useful artifact reduction than purely algorithmic methods.

Our results demonstrate that integrating radiologist feedback into the self-training loop leads to substantial improvements in both quantitative metrics and clinical utility. The framework's strong performance across different domains—from synthetic test cases to real clinical data from different institutions and anatomical regions—confirms the value of this expert-guided approach in bridging the domain gap that has limited previous MAR methods.

Beyond its specific application to metal artifact reduction, RISE-MAR suggests a broader paradigm for developing AI systems that incorporate human judgment not merely as validation but as an essential component of the learning process itself. This human-in-the-loop approach recognizes that in domains where interpretation matters—whether medical images, literary texts, or other complex cultural artifacts—meaning emerges not from data alone but through its engagement with qualified readers within interpretive communities.

As AI continues to advance in healthcare and other interpretive fields, approaches that preserve this dialectical relationship between computational analysis and human judgment will be essential for developing systems that augment rather than replace human expertise. The future of AI in such domains lies not in autonomous interpretation but in collaborative systems that combine the pattern-recognition capabilities of machine learning with the contextual understanding and ethical judgment of human experts.

References:

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In Proceedings of the International Joint Conference on Neural Networks (IJCNN) (pp. 1-8).

2. Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning (pp. 214-223).

3. Choi, Y.-J., Kwon, D., & Baek, S. J. (2024). Dual domain diffusion guidance for 3D CBCT metal artifact reduction. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (pp. 7950-7959).

4. Du, M., Liang, K., Liu, Y., & Xing, Y. (2021). Investigation of domain gap problem in several deep-learning-based CT metal artefact reduction methods. arXiv:2111.12983.

5. Du, M., Liang, K., Zhang, L., Gao, H., Liu, Y., & Xing, Y. (2023). Deep-learning-based metal artefact reduction with unsupervised domain adaptation regularization for practical CT images. IEEE Transactions on Medical Imaging, 42(8), 2133-2145.

6. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(1), 1-35.

7. Gjesteby, L., De Man, B., Jin, Y., Paganetti, H., Verburg, J., Giantsoudi, D., & Wang, G. (2016). Metal artifact reduction in CT: Where are we after four decades? IEEE Access, 4, 5826-5849.

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

9. Kilby, W., Sage, J., & Rabett, V. (2002). Tolerance levels for quality assurance of electron density values generated from CT in radiotherapy treatment planning. Physics in Medicine & Biology, 47(9), 1485-1492.

10. Li, Z., Ma, C., Chen, J., Zhang, J., & Shan, H. (2023). Learning to distill global representation for sparse-view CT. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 21139-21150).

11. Liao, H., Lin, W.-A., Zhou, S. K., & Luo, J. (2019). ADN: Artifact disentanglement network for unsupervised metal artifact reduction. IEEE Transactions on Medical Imaging, 39(3), 634-643.
12. Lin, W.-A., Liao, H., Peng, C., Sun, X., Zhang, J., Luo, J., Chellappa, R., & Zhou, S. K. (2019). DuDoNet: Dual domain network for CT metal artifact reduction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10512-10521).
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9992-10002).
14. Lyu, Y., Fu, J., Peng, C., & Zhou, S. K. (2021). U-DuDoNet: Unpaired dual-domain network for CT metal artifact reduction. In Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (pp. 296-306).
15. Ma, C., Li, Z., He, J., Zhang, J., Zhang, Y., & Shan, H. (2023). Prompted contextual transformer for incomplete-view CT reconstruction. arXiv:2312.07846.
16. Meyer, E., Raupach, R., Lell, M., Schmidt, B., & Kachelrieß, M. (2010). Normalized metal artifact reduction (NMAR) in computed tomography. Medical Physics, 37(10), 5482-5493.
17. Wang, T., Lei, Y., Tang, H., He, Z., Castillo, R., Wang, C., Li, D., Higgins, K., Liu, T., Curran, W. J., Zhou, W., & Yang, X. (2023). SemiMAR: Semi-supervised learning for CT metal artifact reduction. IEEE Journal of Biomedical and Health Informatics, 27(11), 5369-5380.
18. Wang, T., Yu, H., Liu, Y., Sun, H., & Zhang, Y. (2023). Building a bridge: Close the domain gap in CT metal artifact reduction. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 206-216).
19. Yan, K., Wang, X., Lu, L., & Summers, R. M. (2018). DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging, 5(3), 036501.
20. Zhang, H., Cissé, M., Dauphin, Y., & López-Paz, D. (2017). Mixup: Beyond empirical risk minimization. In Proceedings of the International Conference on Learning Representations (pp. 1-12).
21. Zhang, Y., & Yu, H. (2018). Convolutional neural network based metal artifact reduction in X-ray computed tomography. IEEE Transactions on Medical Imaging, 37(6), 1370-1381.