Zero-Shot Super-Resolution for Low-Dose CBCT Images Using Lightweight StereoMamba

Simin Mirzaei, Zhenchao Ma, Hamid Reza Tohidypour, and Panos Nasiopoulos
Electrical & Computer Engineering, University of British Columbia
Vancouver, BC, Canada
Email: {siminmirzaei, zhenchaoma, htohidyp, panosn}@ece.ubc.ca

***Abstract*—Cone-Beam Computed Tomography (CBCT) is a crucial imaging tool in medical diagnostics, but low-dose scans—necessary for minimizing patient radiation exposure—often suffer from degraded spatial resolution. Enhancing the visual quality of these scans is essential for accurate diagnosis and treatment planning. This paper introduces two primary contributions to address this challenge. First, we systematically evaluate the effectiveness of state-of-the-art super-resolution techniques on low-dose CBCT images. Due to the scarcity of real CBCT datasets, which are limited by radiation exposure constraints, we explore the potential of pre-trained stereo super-resolution models originally developed for RGB images. Unlike traditional CBCT datasets that rely on artificially synthesized image pairs, we employ a zero-shot approach to assess the adaptability of these pre-trained models to CBCT imaging. Second, our analysis reveals that existing deep-learning-based super-resolution networks struggle to generalize effectively to CBCT data. To address this, we develop a lightweight adaptation of StereoMamba, a model optimized for natural images, and tailor it to the structural characteristics of CBCT scans. Without requiring additional training, our optimized network achieves the highest Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) scores among all tested models, significantly enhancing CBCT image quality. Our approach makes a significant contribution to improving the quality of low-dose CBCT imaging and provides a path forward for improving diagnostic accuracy and clinical outcomes without increasing radiation risk.**

Keywords—Cone-Beam Computed Tomography; spatial resolution; StereoMamba; lightweight models; zero-shot learning.

I. INTRODUCTION

Cone-Beam Computed Tomography (CBCT) is a crucial imaging technique extensively utilized in dentistry and medical diagnostics, offering detailed 3D visualization of anatomical structures [1][2]. In dental practice, CBCT has revolutionized diagnosis and treatment planning by providing accurate imaging of the maxillofacial region, enabling clinicians to easily examine axial, sagittal, coronal, and custom plane sections [3]. However, achieving high-resolution imaging necessitates increased radiation exposure, heightening the risk of radiation-induced cancers [4]. To mitigate this issue, low-dose CBCT imaging has been developed, which leads to fewer X-ray photons reaching the detector, resulting in lower Signal-to-Noise Ratio (SNR) and reduced spatial resolution [5][6].

Numerous studies have investigated super-resolution, primarily using Single-Image Super-Resolution (SISR) techniques, across various applications. To some extent, these methods can also be applied to CBCT imaging [7][8][9][10][11]. For instance, Hwang et al. [7] investigated the use of a Very Deep Super-Resolution (VDSR) network to restore compressed CBCT images, demonstrating a significant improvement in image quality. Their findings highlight the potential of this approach for clinical applications, enabling reduced storage requirements while maintaining diagnostic accuracy. In another study, Oyama et al. [8] presented a method to suppress artifacts in CBCT images by combining low-resolution images with corresponding high-resolution ones through super-resolution techniques, thereby enhancing clarity and reducing artifacts. Liu et al. [9] uses a Generative Adversarial Network (GAN) to enhance the resolution of low-dose CT images by employing a pyramidal attention model and multiple residual dense blocks to focus on high-frequency image information. Shen et al. [10] proposed a new super-resolution network that uses deep gradient information to guide the reconstruction of CT images, merging gradient image features into the super-resolution branch to enhance structural preservation and detail restoration, ultimately improving image quality. Saharia et al. [11] introduced SR3, a method that uses denoising diffusion probabilistic models to enhance image resolution. Moreover, Liang et al. [12] proposed a hierarchical transformer-based framework for image super-resolution that utilizes the Swin Transformer's efficient window-based self-attention mechanism to effectively capture both local and global dependencies, achieving state-of-the-art performance and surpassing traditional Convolutional Neural Network (CNN)-based methods in both quality across diverse benchmarks. These existing techniques primarily rely on single-image information to enhance the resolution of individual images.

Recognizing this limitation, research on image super-resolution has expanded beyond single-image techniques, with numerous studies exploring approaches that incorporate multiple dependent input images, such as Stereo Super-Resolution (SSR) techniques. These approaches leverage shared information between the input images, resulting in higher spatial resolution for each image [13][14][15][16][17][18]. For example, Lu et al. [13] proposed a two-stage Cycle-consistency network which reconstructs thin-slice MR images by leveraging information from adjacent thick slices to enhance resolution and ensure consistency. In another study, Zhao et al. [14] proposed a self-supervised deep learning algorithm for MRI applications that utilizes shared data between adjacent slices to enhance MRI resolution. By learning spatial correlations across neighboring slices, this method improves both resolution and image quality. Sood et al. [15] presented a method for accurately aligning pre-surgical prostate MRI with histopathology images using super-resolution volume reconstruction. They developed a multi-image super-resolution GAN that leverages multiple lower-resolution MRI images as input to generate high-resolution 3D MRI volumes. Cai et al. [16] proposed a deep learning framework that enhances the segmentation of hepatic ducts in

 ISBN: 978-1-68558-270-8

CT scans. To achieve high-quality segmentation, their method includes an inter-slice super-resolution subnetwork which uses information from neighboring slices in a CT volume to generate intermediate slices, effectively increasing the resolution by interpolating between the existing CT slices. Stereo image super-resolution, which focuses on reconstructing high-resolution details from low-resolution left and right image pairs, has gained significant attention in recent years [17][18]. For example, Chu et al. [17] proposed a novel approach for stereo image super-resolution by extending the NAFNet architecture with components that effectively leverage cross-view information between left and right images. This design enhances spatial coherence and depth consistency, achieving state-of-the-art performance on stereo image super-resolution benchmarks with superior visual quality compared to single-image super-resolution methods. Recently, Ma et al. [18] proposed StereoMamba, a stereo image super-resolution method built on Structured State Space Models (SSMs). The method leverages the Mamba architecture to effectively capture inter-view correlations between left and right image pairs and a Stereo Bi-directional Cross-Attention Module (SBCAM) to further enhance stereo view coherence. Experimental results showed that StereoMamba outperforms state-of-the-art methods for both single and stereo image super-resolution tasks.

In this paper, we present two key contributions. First, we evaluate the performance of state-of-the-art super-resolution methods on low-dose CBCT scans to assess their effectiveness for CBCT applications. Given the scarcity of real CBCT datasets due to radiation exposure constraints, we focus on leveraging pre-trained deep learning models for stereo super-resolution. Since existing CBCT datasets primarily consist of artificially generated pairs with limited practical utility, we adopt a zero-shot strategy, utilizing models pre-trained on large RGB datasets to examine their applicability and generalizability to CBCT imaging.

Second, based on our findings that complex pre-trained super-resolution networks struggle to generalize effectively to CBCT images, we develop a lightweight version of StereoMamba, the network shown to excel on natural images. By adapting this network to the simpler structural patterns of CBCT images, our lightweight model performs exceptionally well on CBCT scans without requiring retraining. This model outperforms all other state-of-the-art super-resolution networks tested, achieving the highest Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) scores, making a significant contribution to improving the quality of low-dose CBCT imaging and providing a path forward for improving diagnostic accuracy and clinical outcomes without increasing radiation risk.

The rest of this paper is organized as follows: Section II presents our objective evaluation of state-of-the-art SISR and SSR methods on CBCT images. Section III details the architecture of our proposed lightweight StereoMamba network, its performance evaluation and discusses the results. Finally, Section V concludes the paper.

II. Performance Evaluation of Super-Resolution Methods for Low-Dose CBCT Scans

In low-dose CBCT imaging, reducing radiation exposure results in lower-resolution scans, making it more challenging for dentists and medical professionals to accurately diagnose conditions. The decrease in X-ray photons reaching the detector lowers the SNR, increases noise, and obscures fine anatomical details, ultimately reducing spatial resolution [6]. This limitation is particularly problematic in dental applications, where high-resolution imaging is essential for precise diagnosis and treatment planning.

We conduct a comprehensive evaluation of super-resolution methods, comparing both SISR and SSR approaches to assess their effectiveness for CBCT imaging. It is worth noting that in general while SISR models perform well on single images, they are inherently limited when compared to SSR methods, as they fail to utilize the redundant spatial information present in adjacent scans. In contrast, SSR methods leverage cross-view correlations, making them more effective for tasks involving medical imaging modalities, such as CBCT. Despite the limitations of SISR approaches, we decided to include the SwinIR network [12] in our evaluation, as it represents the most advanced SISR method available.

Theoretically, deep learning-based super-resolution approaches can enhance low-dose CBCT scans, but their effectiveness is hindered by the scarcity of high-quality training datasets. Unlike natural image super-resolution, which benefits from large datasets of paired low- and high-resolution images, CBCT imaging faces significant constraints. Acquiring both low-dose and high-dose scans from the same patient is not only impractical but also ethically impossible due to radiation exposure risks. As a result, existing CBCT datasets are small and often rely on synthetically generated high-resolution images, which may introduce biases and fail to accurately represent real-world conditions. This data limitation makes it difficult for deep learning models to generalize effectively, often leading to overfitting and reducing their practical applicability for improving CBCT image quality.

To overcome these challenges, we avoid retraining state-of-the-art super-resolution networks used in our evaluation and instead we use the corresponding pre-trained deep learning models. For the case of stereo super-resolution models, given the structural and spatial correlation between adjacent CBCT slices, we treat them as stereo image pairs, allowing us to exploit inter-slice redundancies to enhance image quality while minimizing radiation exposure. Figure 1 illustrates a sequence of transverse slices from a CBCT volume, highlighting the structural continuity between adjacent scans.

For our evaluation of SSR approaches, we selected the top-performing models, specifically variants of NAFSSR [17], along with StereoMamba [18]. The former include NAFSSR-T, the smallest version with minimal parameters; NAFSSR-S, an optimized variant balancing efficiency and performance; NAFSSR-B, a balanced model that maintains high reconstruction quality while reducing computational complexity; and NAFSSR-L, the largest variant designed for maximal reconstruction fidelity. Additionally, we included StereoMamba, which has been shown to outperform all NAFSSR variants on natural images.

By adopting this zero-shot learning strategy, we assess the generalizability of both SISR and SSR pre-trained models to CBCT imaging without additional fine-tuning, demonstrating their potential to improve resolution despite being trained on natural images.

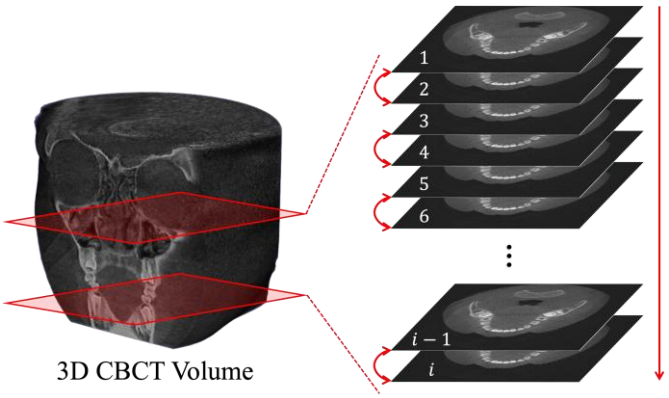

Figure 1. A series of 2D slices from the transverse view in a 3D CBCT volume.

A. *Results of Our Comparative Evaluation*

We use single CBCT slices as input for the SISR model (SwinIR) and adjacent CBCT slices as input pairs for SSR models (NAFSSR and StereoMamba), treating them as two-view images to capture inter-slice dependencies. The evaluation results are summarized in Table I. We observe that all tested networks achieve similar PSNR and SSIM values, with StereoMamba performing the worst. This is unexpected, as StereoMamba has demonstrated superior performance on natural images [18]. These results suggest that although StereoMamba has the most promising architecture, it does not generalize well to CBCT images. This anomaly is due to the fact that grayscale CBCT images tend to have simpler texture and structures and lack color channels, forcing larger, more complex models to overfit to dataset-specific details, amplifying noise or irrelevant features and resulting in suboptimal performance.

Motivated by this finding, we decided to develop a lightweight version of the StereoMamba network, specifically designed to better adapt to the characteristics of CBCT images. The details of this optimized model and its performance are presented in the following section.

III. Overall Architecture of Our Lightweight StereoMamba and Performance Evaluation

In this contribution, we modified the original StereoMamba network presented in [18] to a lightweight version, as illustrated in Figure 2. The original StereoMamba model, while effective for natural images, struggled to generalize to grayscale CBCT images due to their simpler textures, fewer structural variations, and lack of color channels. To enhance generalization, we reduced the number of Residual State Space Groups (RSSGs) to four and the number of SBCAMs to three, as shown in Figure 2. Key differences between our lightweight StereoMamba and the original model include a reduction in the embedding dimension—from 120 to 60—in both the shallow and deep feature extraction stages, as well as a decrease in the expansion ratio of the Vision State Space Module (VSSM) from 2 to 1.2 [18][19]. The process starts with two consecutive low-resolution CBCT slices, labeled as slices $i-1$ and i, serving as inputs. Each pair of neighboring slices is processed sequentially, ensuring no overlap between the groups. The slices first pass through a 3×3 convolutional layer to extract fundamental features, capturing key patterns necessary for subsequent processing. The extracted features are then fed into the deep feature extraction section, which employs four RSSGs connected through residual links. These modules process input features through linear transformations, depth-wise convolutions, and SiLU activation to capture both local and global spatial correlations. Residual connections are included to preserve input information and improve feature propagation. This design ensures feature continuity, enhances gradient flow during training, and refines feature representations. To strengthen the interaction between adjacent slices $i-1$ and i, the deep feature extraction section incorporates SBCAM modules. These modules employ bi-directional cross-attention mechanisms to enable a synergistic exchange of contextual information between adjacent CBCT slices, improving the network's ability to capture inter-slice relationships. Finally, the reconstruction module processes the refined feature maps from the deep feature extraction section, integrating them to generate high-resolution reconstructions of input adjacent CBCT slices.

Our proposed complexity reduction allows this lightweight version to avoid memorizing dataset-specific noise and instead to focus on essential structural details, leading to more robust feature extraction. Additionally, fewer layers reduce the accumulation of redundant or noisy representations, mitigating the risk of amplifying artifacts. In summary, this streamlined architecture aims to enhance generalization without retraining and achieve high-quality super-resolution results, making it well-suited for CBCT applications.

Table II shows the performance of our pretrained lightweight StereoMamba model, as well as that of the state-of-the-art networks included in Table I in terms of PSNR and SSIM. The evaluation results demonstrate that our lightweight StereoMamba model outperforms all other methods, achieving superior image quality in terms of both structural preservation and perceptual similarity. Notably, NAFSSR-T, which also has a relatively small number of parameters, also performs well, reinforcing the idea that lighter models are more robust to domain shifts from RGB to grayscale CBCT images. These models avoid overfitting by focusing on fundamental spatial features rather than dataset-specific variations. In summary, our evaluation demonstrates that our lightweight StereoMamba model effectively balances complexity and performance, achieving state-of-the-art results in CBCT image super-resolution while maintaining computational efficiency. Its ability to generalize across different datasets, combined with reduced inference time, makes it a strong candidate for practical applications in dental and medical imaging.

TABLE I. Performance Comparison of State-of-the-Art SISR and SSR Methods for CBCT Resolution Enhancement (#S and #Params Indicate the Scale Factor and the Number of Parameters, Respectively).

Method	#S	#Params	PSNR	SSIM
SwinIR	x2	11.28M	36.32	0.9849
NAFSSR-T	x2	0.45M	36.72	0.9865
NAFSSR-S	x2	1.54M	36.65	0.9859
NAFSSR-B	x2	6.77M	36.47	0.9857
NAFSSR-L	x2	23.79M	36.08	0.9846
StereoMamba	x2	7.55M	35.88	0.9840

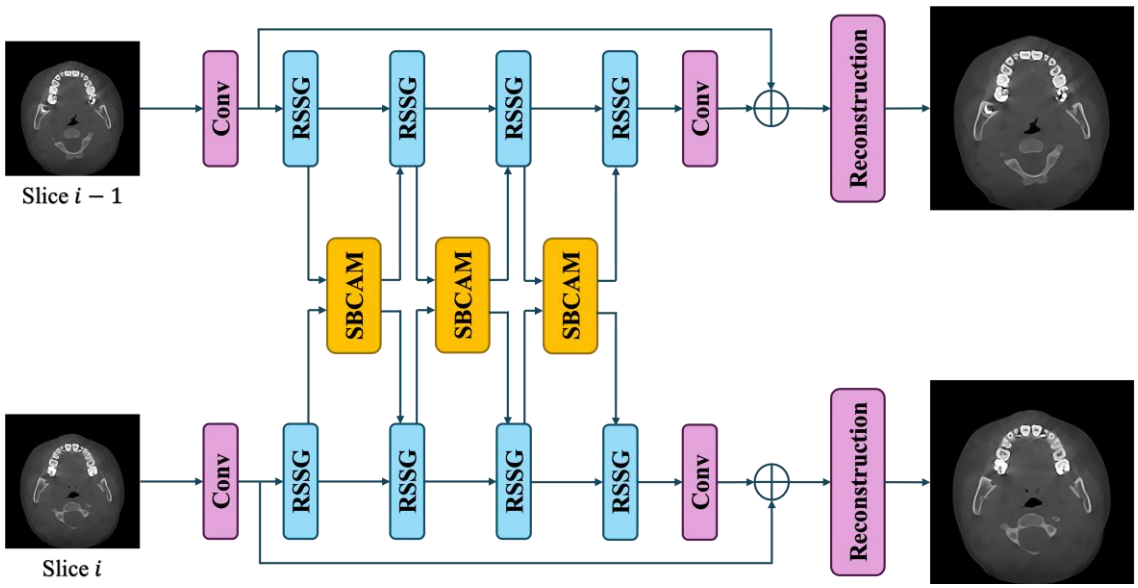

Figure 2. Architecture of our lightweight StereoMamba network for enhancing the resolution of adjacent slices in a 3D CBCT volume.

TABLE II. PERFORMANCE COMPARISON OF OUR LIGHTWEIGHT STEREOMAMBA AGAINST THE STATE-OF-THE-ART SUPER-RESOLUTION METHODS (#S AND #PARAMS INDICATE THE SCALE FACTOR AND THE NUMBER OF PARAMETERS, RESPECTIVELY).

Method	#S	#Params	PSNR	SSIM
SwinIR	x2	11.28M	36.32	0.9849
NAFSSR-T	x2	0.45M	36.72	0.9865
NAFSSR-S	x2	1.54M	36.65	0.9859
NAFSSR-B	x2	6.77M	36.47	0.9857
NAFSSR-L	x2	23.79M	36.08	0.9846
StereoMamba	x2	7.55M	35.88	0.9840
StereoMamba-Light	x2	**0.9M**	**38.03**	**0.9886**

IV. CONCLUSION AND FUTURE WORK

In this paper, we conducted a comprehensive evaluation of super-resolution techniques for low-dose CBCT imaging, comparing state-of-the-art SISR and SSR methods. A key challenge in applying deep learning-based super-resolution to CBCT imaging is the scarcity of high-quality paired training datasets. To mitigate this limitation, we adopted a zero-shot learning approach, utilizing pre-trained networks without additional fine-tuning. Our evaluation demonstrated that the original StereoMamba model, despite its success in natural image super-resolution, struggled to generalize to grayscale CBCT images. Motivated by this finding, we developed a lightweight version of StereoMamba, reducing model complexity while preserving essential spatial features. The results indicate that our proposed lightweight StereoMamba model outperforms existing SISR and SSR methods, achieving improved spatial resolution and structural preservation in low-dose CBCT scans. By reducing network complexity and focusing on fundamental spatial correlations, our lightweight StereoMamba provides a practical and effective solution for enhancing CBCT image quality without requiring extensive retraining. This advancement has significant implications for dental and medical applications, where improved CBCT resolution can enhance diagnostic accuracy while minimizing radiation exposure. Future work will explore domain-specific fine-tuning strategies and further optimization of SSR architectures to enhance generalizability across different medical imaging modalities.

REFERENCES

[1] B. Schlueter, K. B. Kim, D. Oliver, and G. Sortiropoulos, “Cone beam computed tomography 3D reconstruction of the mandibular condyle,” The Angle Orthodontist, vol. 78, no. 5, pp. 880–888, 2008.

[2] W. De Vos, J. Casselman, and G. R. J. Swennen, “Cone-beam computerized tomography (CBCT) imaging of the oral and maxillofacial region: A systematic review of the literature,” International journal of oral and maxillofacial surgery, vol. 38, no. 6, pp. 609–625, 2009.

[3] C. Angelopoulos, “Anatomy of the maxillofacial region in the three planes of section,” Dental Clinics, vol. 58, no. 3, pp. 497-521, 2014.

[4] J. B. Carter, J. D. Stone, R. S. Clark, and J. E. Mercer, “Applications of cone-beam computed tomography in oral and maxillofacial surgery: an overview of published indications and clinical usage in united states academic centers and oral and maxillofacial surgery practices,” Journal of Oral and Maxillofacial Surgery, vol. 74, no. 4, pp. 668–679, 2016.

[5] S. Kawauchi, K. Chida, Y. Hamada, and W. Tsuruta, “Image quality and radiation dose of conventional and wide-field high-resolution cone-beam computed tomography for cerebral angiography: a phantom study,” Tomography, vol. 9, no. 5, pp. 1683-1693, 2023.

[6] L. Yu et al., “Dose and image quality evaluation of a dedicated cone-beam CT system for high-contrast neurologic applications,” American Journal of Roentgenology, vol. 194, no. 2, pp. W193-W201, 2010.

[7] J. J. Hwang, Y. H. Jung, B. H. Cho, and M. S. Heo, “Very deep super-resolution for efficient cone-beam computed tomographic image restoration,” Imaging Science in Dentistry, vol. 50, no. 4, p. 331, 2020.

[8] A. Oyama et al., “Image quality improvement in cone-beam CT using the super-resolution technique,” Journal of radiation research, vol. 59, no. 4, pp. 501-510, 2018.

[9] X. Liu et al., “Super-resolution reconstruction of ct images based on multi-scale information fused generative adversarial networks,” Annals of Biomedical Engineering, vol. 52, no. 1, pp. 57-70, 2024.

[10] Y. Shen et al., “Ct image super-resolution under the guidance of deep gradient information,” Journal of X-Ray Science and Technology, vol. 33, no. 1, pp. 58-71, 2025.

[11] C. Saharia et al., "Image super-resolution via iterative refinement," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 4, pp. 4713-4726, 2022.

[12] J. Liang et al., "Swinir: Image restoration using swin transformer," Proceedings of the IEEE/CVF international conference on computer vision, pp. 1833-1844, 2021.

[13] Z. Lu, Z. Li, J. Wang, J. Shi, and D. Shen, "Two-stage self-supervised cycle-consistency network for reconstruction of thin-slice MR images," Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, pp. 3-12, 2021.

[14] C. Zhao et al., "SMORE: a self-supervised anti-aliasing and super-resolution algorithm for MRI using deep learning," IEEE transactions on medical imaging, vol. 40, no. 3, pp. 805-817, 2020.

[15] R. R. Sood et al., "3D Registration of pre-surgical prostate MRI and histopathology images via super-resolution volume reconstruction," Medical image analysis, vol. 69, p. 101957, 2021.

[16] N. Cai et al., "W-Shaped Net: An Inter-Slice Super-Resolution Segmentation Deep Network for CT Scans of Hepatic Ducts," Electronics, vol. 14, no. 2, p. 321, 2025.

[17] X. Chu, L. Chen, and W. Yu, "Nafssr: Stereo image super-resolution using nafnet," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1239-1248, 2022.

[18] Z. Ma, H. R. Tohidypour, P. Nasiopoulos, and V. C. M. Leung, "StereoMamba: Enhancing Stereo Image Super-Resolution with Structured State Space Models and Bi-Directional Cross Attention," ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5, 2025.

[19] H. Guo et al., "Mambair: A simple baseline for image restoration with state-space model," European conference on computer vision, pp. 222-241, 2024.